Investigating the genetic basis of reading and language skills

Proefschrift
ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. Th.L.M. Engelen,
volgens besluit van het college van decanen
in het openbaar te verdedigen op vrijdag 11 september 2015
om 10.30 uur precies

door
Alessandro Gialluisi
geboren op 26 augustus 1984
te Castellana Grotte (Italië)

Promotor

Prof. dr. Simon E. Fisher

Copromotor

Dr. Clyde Francks (MPI)

Manuscriptcommissie

Prof. dr. Barbara Franke

Prof. dr. Ben A. M. Maassen (Rijksuniversiteit Groningen)

Prof. dr. Thomas Bourgeron (Université Paris Diderot et Institut Pasteur, Frankrijk)

Investigating the genetic basis of reading and language skills

Doctoral Thesis
to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. Th.L.M. Engelen,
according to the decision of the Council of Deans
to be defended in public on Friday, September 11, 2015
at 10.30 hours

by
Alessandro Gialluisi
born on August 26, 1984
in Castellana Grotte (Italy)

Supervisor

Prof. dr. Simon E. Fisher

Co-supervisor

Dr. Clyde Francks (MPI)

Doctoral Thesis Committee

Prof. dr. Barbara Franke

Prof. dr. Ben A. M. Maassen (University of Groningen)

Prof. dr. Thomas Bourgeron (Université Paris Diderot et Institut Pasteur, France)

"There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved."

Charles Robert Darwin,

On the Origin of Species by Means of Natural Selection

© 2015 Alessandro Gialluisi

All rights reserved. No part of this thesis may be reproduced or printed in any form, by any electronic or mechanical means, without written permission of the author.

Cover design: Mauro Giannoccaro

Printed by: Ipskamp Drukkers, Enschede

Contents

Chapter 1:
General introduction

General introduction

Reading and language capacities represent fundamental keystones in the developmental route of a child, and impairment of these abilities can have long-lasting effects, resulting in relatively reduced educational and professional achievements and low socioeconomic status in adulthood (Bishop & Snowling, 2004; Pennington & Bishop, 2009). In this chapter I provide an overview of the most frequently studied reading and language skills along with two of the most prevalent and investigated language-related disorders, namely Reading Disability and Specific Language Impairment. A complete view of these disorders will allow us to understand more of the neuropsychology, neurobiology and genetics supporting reading and language skills, which represent the main object of investigation of the present thesis.

How is reading/language performance assessed? The use of continuous traits

Although the concept of reading and language performance may appear quite simple to understand, defining in an objective way how "well" an individual can read or speak is far from easy. Indeed, assessing reading and language skills not only means testing the actual capacity of a subject to read (usually represented by word reading, spelling and reading comprehension) and to speak/understand oral language (usually assessed through expressive/receptive language scores). It also means assessing various cognitive skills underlying written and oral language capacities, such as phoneme awareness and phonological short term memory (see Table 1 for a complete list and definition of these traits), which are often handicapped in subjects with poor reading/language performance and represent part of their core cognitive deficits. Several psychometric tests, each measuring a specific reading-/language-related trait, have been developed for this purpose. These continuous traits tap into diverse cognitive domains underlying reading and language, generally show strong intercorrelations -underpinned by common environmental and genetic influences (Harlaar et al., 2008; Logan et al., 2011)- and tend to have a normal distribution in the general population (Figure 1). In this view, Reading Disability (RD, also known as developmental dyslexia) and Specific Language Impairment (SLI, sometimes referred to as Language Impairment, LI) represent the lower tails of these distributions, and can therefore help us to understand more of the neuropsychological, neurobiological and genetic basis of reading and language (reviewed below).

Trait	Description (ability assessed)
Word Reading[a]	Reading real words
Word Spelling[a]	Spelling real words
Reading Comprehension[a]	Ability to read text, process it and understand its meaning
Phonological Decoding	Ability to convert letter strings into sounds, according to given phonetic rules
Phoneme Awareness	Ability to recognize and manipulate speech sounds (phonemes)
Orthographic Coding	Ability to recognize a word as an orthographic unit and to retrieve the corresponding phonological form
Rapid Automatized Naming	Ability to rapidly produce verbal labels for visual stimuli (colors, numbers, letters, pictures)
Processing Speed	Ability to automatically and fluently perform relatively easy or over-learned cognitive tasks
Nonword repetition[b]	Ability to repeat nonsense words orally presented (phonological short term memory)
Expressive Language[b]	Sentence recalling and production (expressive domain of language)
Receptive Language[b]	Listening and auditory comprehension (receptive domain of language)

Table 1. Cognitive traits routinely used to assess reading and language performance. [a] Commonly used to diagnose RD. [b] Commonly used to diagnose SLI.

Reading Disability (RD) and Specific Language Impairment (SLI): a brief overview

Definition and Diagnosis

RD is defined as a difficulty/delay in the acquisition of written language ability that cannot be explained by obvious causes, such as low IQ, sensory impairments or lack of educational opportunity; while SLI is defined as an unexpected difficulty/delay in acquiring oral language abilities, despite normal hearing and intelligence, and in the absence of overt neurological deficits or other syndromes (American Psychiatric Association, 2013).

These disorders are usually diagnosed when a subject shows typical reading or language scores at least 1.5 standard deviations (SDs) below the normative mean of the general population, matched for age and IQ (Peterson and Pennington, 2012; Reader et al., 2014). Nonetheless, this cutoff threshold is somewhat arbitrary (Pennington & Bishop, 2009; Peterson and Pennington, 2012; Raskind et al., 2013) and may vary across studies, usually ranging between -2.0 and -1.0 SDs (Bishop & Snowling, 2004; Newbury et al., 2014; Newbury et al., 2010; Willcutt et al., 2005; Willcutt et al., 2010). This partly explains the variation in the epidemiological estimates of these disorders (see paragraph below and Table 2). Traits routinely assessed to diagnose RD include word reading, spelling and reading

comprehension, while expressive and receptive language and nonword repetition are commonly tested to diagnose SLI.

Another matter of debate in RD and SLI definition is the role of general cognitive abilities in reading and language capacities. The diagnosis of RD and SLI has been customarily based on the discrepancy between poor reading/language performance and normal general intelligence, and low IQ was unanimously considered an exclusion criterion for the diagnosis of these disorders (Bishop & Snowling, 2004; Pennington & Bishop, 2009). More recently this approach has been criticized, as poor readers/speakers with normal nonverbal IQ often show the same underlying deficits as cases with low nonverbal IQ (see *Neuropsychology* paragraph below; Bishop & Snowling, 2004; Pennington & Bishop, 2009). It this view, it would be more appropriate to speak of Language Impairment (LI, without the "Specific" prefix) rather than SLI, and RD and LI can be more broadly conceived as disorders characterized by poor reading and language performance, regardless of the general intelligence of subjects (Bishop & Snowling, 2004; Pennington & Bishop, 2009). However, part of the scientific community keeps on considering the role of IQ-based discrepancy in the definition of reading and language impairments (American Psychiatric Association, 2013; Peterson & Pennington, 2012; Newbury et al., 2014; Raskind et al., 2013; Reader et al., 2014). For simplicity, in this thesis I will use the traditional terms RD and SLI, meaning a broad category of reading and language impairments. Also, I will have a neutral approach towards the role of nonverbal intelligence in reading and language cognition, analyzing reading-language traits both before and after adjustment for performance IQ (see *Aims of this thesis* paragraph below).

Epidemiology

From an epidemiological point of view, RD and SLI show similar characteristics (summarized in Table 2), which reflect the strong intercorrelations between reading and language skills. RD has a prevalence of 5-10% among school-aged children in many populations (Pennington, 1990; Shaywitz et al., 1990), with males being more frequently affected than females (sex ratio M:F between 1.9 and 3.3 in epidemiological samples, Rutter et al., 2004). Similarly, SLI is relatively frequent among school-aged children, with a prevalence of 5-8% in English-speaking populations (Tomblin et al., 1997; Law et al., 1998), and is more prevalent in males than in females (sex ratio 1.5 in epidemiological sample; Tomblin et al., 1997). Nonetheless, it necessarily has an earlier onset compared to RD, as

children need to acquire spoken language abilities before reading skills (Snowling et al., 2000; Bishop & Snowling, 2004; Harlaar et al., 2008). The variability in epidemiological estimates can be explained by several factors, including different inclusion/exclusion criteria and diagnostic cutoff thresholds (as already discussed above).

RD and SLI also show notable phenotypic and clinical overlaps: 43% of SLI children are later diagnosed with RD (Snowling et al. 2000) and up to 55% of dyslexic children meet criteria for SLI (McArthur et al. 2000). Even when RD children do not meet criteria for SLI diagnosis, they often present milder forms of language delays (Bishop and Snowling, 2004), and children with SLI are much more likely to develop reading difficulties than children with normal language abilities (Catts et al., 2002). Moreover, RD and SLI frequently co-occur with other neurodevelopmental disorders, such as Attention Deficit Hyperactivity Disorder (ADHD) (Wilcutt et al. 2010; Pennington, 2006; Mueller, 2012), and Speech Sound Disorder (SSD), another typical speech and language delay (Pennington & Bishop, 2009). Indeed, RD shows comorbidity rates of ~25-40% with ADHD (Wilcutt et al. 2010) and ~10-30% with SSD (Pennington & Bishop, 2009); while SLI studies have reported comorbidities of ~20-90% with ADHD (Mueller, 2012) and ~5-15% with SSD (Shriberg et al., 1999). These data suggest the presence of shared neurobiological deficits underlying these disorders. This hypothesis is supported also at the neuropsychological level, where some cognitive deficits appear to be involved both in RD and in SLI psychopathology (see *Neuropsychology* paragraph below).

Disorder	Exclusion criteria	Prevalence	Heritability	Sex ratio (M:F)	Comorbidities
RD	Neurological deficits/brain damage Inadequate intelligence Lack of educational opportunity Hearing/visual impairments	5-10%[a]	40-60%[b]	1.9-3.3[c]	43-55% with SLI[d] 25-40% with ADHD[e] 10-30% with SSD[f]
SLI	Neurological deficits/brain damage Inadequate intelligence Hearing impairments Known syndromes (e.g. autism)	5-8%[g]	50-75%[h]	1.5[i]	43-55% with RD[d] 20-90% with ADHD[j] 5-15% with SSD[k]

Table 2. Epidemiology of RD and SLI. Note: prevalence and comorbidity estimates vary due to different cutoff values used for diagnosis. Sex ratios refer to epidemiological samples. [a] Pennington, 1990; Shaywitz et al., 1990. [b] Raskind et al., 2013; [c] Rutter et al., 2004; [d] Snowling et al. 2000; McArthur et al. 2000; [e] Wilcutt et al. 2010; [f] Peterson et al., 2009; [g] Tomblin et al., 1997; Law et al., 1998; [h] Bishop & Hayiou-Thomas; [i] Tomblin et al., 1997; [j] Mueller, 2012; [k] Shriberg, 1999.

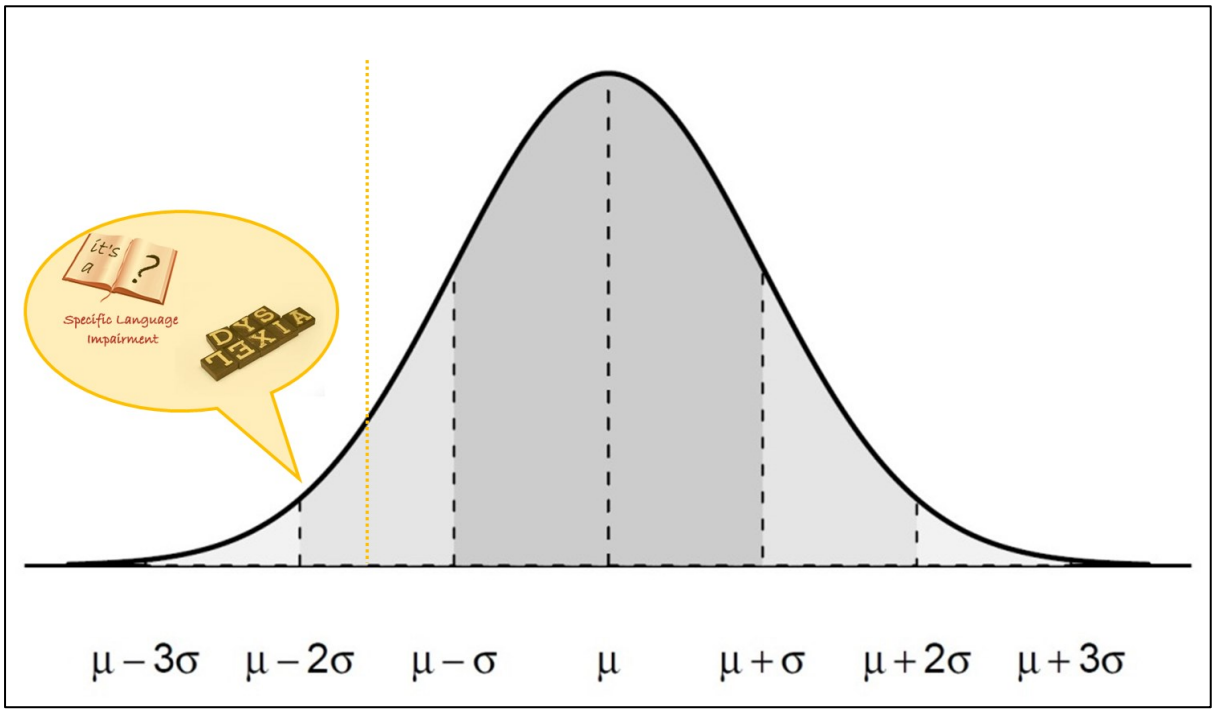


Figure 1. Abstract representation of the epidemiology of RD and SLI. Reading- and language- related skills can be viewed as continuous traits which tend to be normally distributed in the general population. In this view, RD and SLI cases constitute the lower tail of this distribution, and are often comorbid. Note: diagnostic cutoff threshold (dashed yellow line) is only indicative.

Neuropsychology

Over the years, several neuropsychological theories have been proposed to elucidate the etiology of RD and SLI (Bishop & Snowling, 2004; Pennington & Bishop, 2009).

A large body of work has shown that many RD cases exhibit impaired phonological processing, (i.e. the ability to process and retrieve speech sounds) (Pennington & Bishop, 2009). This impairment -often reflected in poor phoneme awareness performance- led to formulation of the phonological theory, which postulates that dyslexia is mainly caused by a deficit in phonological representations, namely in converting graphemes (combinations of letters that are pronounced together, as a unit) to phonemes (the smallest phonetic units of speech sound, resulting from reading graphemes) (Paracchini et al., 2007). Other theories have proposed a role in RD etiology for many other cognitive processes, including impaired visual/auditory sensory modalities and fine motor control (which may be integrated within the magnocellular theory; Stein et al., 2001), poor orthographic coding, rapid automatized

naming and processing speed (see Bishop and Snowling, 2004; Pennington & Bishop, 2009 for a review).

Similarly, a number of theories have been advanced to explain SLI. For example, the rapid temporal processing theory postulates that poor temporal resolution of auditory perceptual systems in affected children may result in altered speech perception and finally in impaired language learning (Tallal, 2004). In other words, this theory -which has also been proposed to explain RD etiology- hypothesizes a role of auditory and phonological processing in mechanisms giving rise to SLI, and the contributions of these two cognitive functions to language performance appear to be largely independent (Bishop et al., 1999). An alternative theory maintains that deficits in phonological short term memory (usually measured through tests of nonword repetition) may account for language impairment, as this cognitive skill is considered to be important not only for learning new vocabulary and syntax, but also for retaining and processing this linguistic knowledge while speaking (Newbury et al., 2005; Pennington & Bishop, 2009). Other deficits detected in language impaired children (reviewed in Bishop and Snowling, 2004; Pennington & Bishop, 2009) affect syntactic skills (e.g. the ability to use the right tense in statements) or procedural learning (i.e. a cognitive process including both short term memory and syntactic skills).

Although none of these theories can explain the totality of RD/SLI cases, we can draw some important conclusions on the neuropsychology of RD and SLI. First, multiple underlying deficits appear to exist at the basis of these disorders, often co-occurring in the same individual, and at least one of these deficits, i.e. phonological processing, is common to both RD and SLI. Second, the most severe cases typically show two or more co-occurring deficits in distinct cognitive skills. Third, cognitive overlaps between RD and SLI and with other neurodevelopmental disorders -such as SSD and ADHD- often reflect clinical overlaps (i.e. comorbidity). This is the case of rapid automatized naming and processing speed deficits, which are detected both in RD and in ADHD (Willcutt et al., 2005; Willcutt et al., 2010; McGrath et al., 2011), and of impaired phonological processing, which is observed not only in RD and SLI, but also in SSD (Peterson et al., 2009; Pennington & Bishop, 2009). In this complex scenario -made up of several links connecting reading, language and other neurodevelopmental cognitive domains- what determines the onset of one disorder rather than another appears to be the kind of specific cognitive deficits that co-occur. As an example, phonological short term memory and/or syntactic deficits often co-occur with phonological deficits in SLI cases (Bishop et al., 2006), while phonological deficits and

impaired rapid automatized naming are detected in RD, but not in SSD (Raitano et al., 2004). Finally, this complex neuropsychological scenario suggests once again the presence of partly shared neurobiological bases between RD and SLI, as well as with other neurodevelopmental disorders such as SSD and ADHD.

Genetic architecture supporting reading and language

Reading and language skills are complex traits, i.e. influenced by a number of environmental and genetic factors, with a substantial genetic influence. Indeed, these skills show moderate to high heritability (representing the fraction of phenotypic variance explained by genetic factors). Independent studies have reported heritability estimates of ~40-70% for several reading and language traits, including word reading, spelling, phoneme awareness, phonological decoding, orthographic coding and nonword repetition (Gayán & Olson, 2001; Francks et al., 2003; Bishop et al., 1999), although some other traits, such as auditory processing in SLI, appear more environmental in origin (Bishop et al., 1999). Comparable estimates have been observed for pairwise bivariate heritabilities among these traits (Gayán & Olson, 2001) and for heritabilities of latent variables underlying reading and language measures (Francks, 2001; Gayán & Olson, 2003; Dale et al., 2010). This lends further support to the presence of a common genetic influence on reading and language skills. Similarly to continuous traits, RD and SLI are etiologically complex phenotypes, which tend to run in families and are moderately heritable (Pennington and Bishop, 2009). Heritability estimates of RD (~40-60%; Raskind et al., 2013; Fisher & Defries, 2002) and SLI (~40-75%; Viding et al., 2004; Bishop & Hayiou-Thomas, 2008) are generally consistent with those observed in continuous traits. Such data (presented in detail in Chapter 2) overall suggest a substantial genetic etiology for these disorders, and strong genetic influences on the underlying reading and language traits. Nonetheless, only a minor part of this heritability has been accounted for by genetic findings (the so-called "missing heritability" issue; Peterson & Pennington 2012; Newbury & Monaco, 2010).

We briefly review below the RD and SLI susceptibility loci/genes identified so far. These genes were mostly identified through linkage analysis, followed by either positional cloning or targeted association mapping, with both categorical RD/SLI and continuous reading/language traits. Several studies have reported associations between reading/language traits and Single Nucleotide Polymorphisms (SNPs, i.e. single-base changes in the genome,

with frequency of at least 1%) both in selected datasets and in general population cohorts (Carrion-Castillo et al., 2013; Reader et al., 2014; see below and Chapter 4 for details), suggesting that these common genetic variants exert their effects across the whole range of distribution of reading and language skills, rather than only on the lower tail of the distribution (i.e. on RD and SLI cases). Moreover, some of these genes (reviewed below) appear to contribute to both reading and language skills (Newbury et al., 2011; Scerri et al., 2011; Bates et al., 2011), indicating that the overlap between these cognitive domains is detected also at the genetic level, and further supporting the hypothesis of partly shared neurobiological mechanisms. Finally, many of these genes play roles in important developmental processes in the Central Nervous System (CNS), such as neuronal migration, axonal guidance, neurite/dendrite outgrowth and synaptic plasticity (see Tables 3 and 4). Therefore, the disruption of molecular pathways underlying these functions has been hypothesized to play a role in RD/SLI etiology and, more in general, to influence reading and language skills (Newbury et al., 2014; Pennington & Bishop, 2009; Peterson & Pennington, 2012; Poelmans et al., 2011).

RD susceptibility loci

Loci frequently linked to RD and/or reading-related traits are shown in Table 3. The most consistent findings have been reported for loci *DYX1, DYX2, DYX3* and *DYX5*, where the involvement of candidate susceptibility genes has been supported by independent studies.

DYX1 (15q21) was the first locus found to be linked to RD, and linkage at this location has been replicated in several studies (Grigorenko et al., 1997; Schulte-Korne et al., 1998; Chapman et al., 2004; Bates et al., 2007; Platko et al., 2008). In this region, *DYX1C1* (*Dyslexia susceptibility 1 candidate 1)* was first identified through a balanced translocation disrupting this gene, which co-segregated with reading difficulties in a Finnish family (Nopola-Hemmi et al., 2000; Taipale et al., 2003). Since then, significant associations with RD and reading-related traits have been reported for many SNPs at this locus. In the original study by Nopola-Hemmi and colleagues (2000), a similar translocation was reported to co-segregate with RD in another dyslexic family. More recently, targeted association analysis of its breakpoint region on chromosome 15 -within the gene *CYP19A1*- revealed moderate SNP associations with categorical dyslexia in several datasets, and with quantitative measures of language and speech, although not always consistently across datasets (Anthoni et al., 2012).

DYX1C1 encodes a protein which appears to be important for neuronal migration, cilia assembly and motility (Tammimies et al., 2013; Chandrasekar et al., 2013; Tarkar et al., 2013). Similarly, the product of *CYP19A1* -the aromatase enzyme, which normally takes part to the conversion of androgens into estrogens- appears to be involved in neuronal migration and dendrite outgrowth processes (Anthoni et al., 2012).

Another well-validated RD susceptibility locus is *DYX2* (6p22.3-p21.3; Cardon et al., 1994; Grigorenko et al., 1997; Fisher et al., 1999; Fisher et al., 2002; Kaplan et al., 2002; Platko et al., 2008), where two candidate genes have been identified. The first one, *KIAA0319*, was identified through candidate SNP association studies, which reported significant associations in the putative promoter region of the gene, both in clinical RD datasets (Francks et al., 2004; Cope et al., 2005; Harold et al., 2006) and in population-based cohorts (Luciano et al., 2007; Paracchini et al., 2008; Scerri et al., 2011). SNPs in *KIAA0319* have also been associated with SLI and continuous language traits (Rice et al., 2009; Newbury et al., 2011), indicating potential pleiotropic effects of this gene on language skills. The second gene in this region, *DCDC2 (doublecortin domain containing 2)*, was first identified through the association of RD and reading-related traits with two genetic variants other than SNPs: a small deletion (Meng et al., 2005; Marino et al., 2012) and a compound Short Tandem Repeat (STR, i.e. a short sequence of DNA that is repeated a variable number of times at a specific location in the genome) within this gene (Schumacher et al., 2006). Additional SNP markers in *DCDC2* have been associated with dyslexia and reading-related traits (Meng et al., 2005; Schumacher et al., 2006; Wilcke et al., 2009; Newbury et al., 2011; Harold et al. 2006; Scerri et al., 2011). Molecular knockdown of Kiaa0319 and Dcdc2 suggest that they are both involved in neuronal migration in the developing CNS (Velayos-Baeza et al., 2007; 2008; Peschansky et al., 2010; Adler et al., 2013; Meng et al., 2005; Wang et al., 2011). Dcdc2 is also thought to have a role in regulating cilia structure, length and signaling (Massinen et al., 2011; Grati et al., 2015; Schueler et al., 2015). However, caution is needed in the interpretation of these studies, especially for Dcdc2, since knockout models of this gene did not show any evidence of neuronal migration deficits, in contrast with knockdown models (Wang et al., 2011). Moreover, off-target effects of short hairpin RNAs (shRNAs), normally used for RNA interference, have been recently reported (Baek et al., 2014).

DYX3 (2p12-p16) was first reported to be linked to RD in a large multigenerational family from Norway (Fagerheim et al., 1999) and later confirmed as a susceptibility locus in several other studies (Petryshen et al., 2002; Francks et al., 2002; Kaminen et al., 2003; Anthoni et

al., 2007; De Kovel et al., 2008). A combined linkage/association study of Finnish RD families refined this finding, reporting significant associations on 2p12, very close to the genes *GCFC2* (*GC-rich sequence DNA-binding factor 2*, also known as *chromosome 2 open reading frame 3,C2ORF3*) and *MRPL19 (mitochondrial ribosomal protein L19*) (Anthoni et al., 2007). This association was internally replicated in an independent sample of dyslexic German families, supporting the hypothesis that variants in these genes may increase susceptibility to RD (Anthoni et al., 2007). Nonetheless, the molecular function of these genes and their role in RD etiology are still unclear.

DYX5 locus (3p12-q13) was first identified in a linkage analysis of a four-generation Finnish family (Nopola-Hemmi et al., 2001) and linkage of this region to RD and reading-related traits was further supported by independent genome-wide scans (Fisher et al., 2002; Bates et al., 2007). The putative causative gene, *ROBO1 (roundabout homolog 1)*, was later discovered in the same pedigree described in the original linkage report, where a rare haplotype associated with reduced gene expression was found to co-segregate with RD (Hannula-Jouppi et al., 2005). Also, *ROBO1* was disrupted by a translocation in an unrelated dyslexic subject (Hannula-Jouppi et al., 2005). More recently, SNP associations with RD and with continuous reading and language traits have been reported in *ROBO1*, both in population-based cohorts (Bates et al., 2011) and in RD datasets (Tran et al., 2014). *ROBO1* encodes an axonal guidance receptor, which drives dendrites in the brain and contributes to several neurodevelopmental processes, including neuronal migration, differentiation and synapse formation (Andrews et al., 2006; 2008).

In addition to these strong candidates, other RD susceptibility loci have been discovered through linkage, although robust candidate genes in these regions -supported by significant genetic associations- have not yet been identified. These loci include *DYX4* (6q11.2-q12; Petryshen et al., 2001; Bates et al., 2007); *DYX6* (18p11.2; Fisher et al., 2002; Bates et al., 2007; Seshadri et al., 2007); *DYX7* (11p15.5 ; Fisher et al., 2002; Hsiung et al., 2004); *DYX8* (1p36-p34; Grigorenko et al., 2001; Tzenova et al., 2004; De Kovel et al., 2008) and *DYX9* (Xq27.3-q28; De Kovel et al., 2004; Bates et al., 2007; Huc-Chabrolle et al., 2013). Similarly, evidence of linkage to RD and/or continuous reading-related traits has been reported by more than one study in additional regions, such as 2q22.3 (Raskind et al., 2005; Bates et al., 2007), 7q32 (Kaminen et al., 2003; Bates et al., 2007), 4q13 (Brkanac et al., 2008; Field et al., 2013), 16p12 and 17q22 (Loo et al., 2004; Field et al., 2013). Nonetheless, causative genetic variants have not yet been identified in these regions.

Locus	Location	Candidate genes[a]	Biological process
DYX1	15q21	DYX1C1	Cilia function and assembly; Neuronal migration
		CYP19	Dendrite outgrowth; Neuronal migration; Steroid hormones metabolism
DYX2	6p22.3-p21.3	DCDC2	Cilia function and assembly; Dendrite outgrowth; Neuronal migration
		KIAA0319	Neuronal migration
DYX3	2p12-p16	MRPL19	unknown
		GCFC2 (C2ORF3)[b]	
DYX4	6q11.2-q12		
DYX5	3p12-q13	ROBO1	Axon guidance
DYX6	18p11.2		
DYX7	11p15.5		
DYX8	1p36-p34		
DYX9	Xq27.3-q28		

Table 3. Loci frequently reported to be linked to RD and reading-related traits. [a] Only candidate genes implicated in RD by genetic associations are reported. [b] Old gene nomenclature as used in the original report (Anthoni et al., 2007) is indicated in brackets.

SLI susceptibility loci

Loci frequently reported to be linked to SLI and/or language-related traits are reported in Table 4. Among these, the involvement of candidate genes in *SLI1* and *SLI4* received strong support from targeted association studies.

SLI1 (16q23.1-q24) has been linked to SLI and language-related traits in various studies (The SLI Consortium, 2002; 2004; Monaco, 2007; Falcaro et al., 2008). Newbury and colleagues (2009) later identified two independent genetic effects in a candidate SNP association analysis of this target region. One was located within *CMIP (c-MAF induced protein)*, encoding an adaptor protein which is hypothesized to act as a cytoskeletal component and to take part to neuronal migration. The other one was located in another gene, *ATP2C2* (*ATPase, Ca^{2+} transporting, type 2C, member 2*), which codes for a calcium transporter ATPase regulating cellular levels of calcium and manganese ions, a key process for synaptic plasticity, transmission and neuronal motility (Newbury & Monaco; 2010). Some of the SNPs associated within *CMIP* were later found to be associated with word reading and spelling,

suggesting pleiotropic effects of this gene on reading and language skills (Newbury et al., 2011; Scerri et al., 2011).

SLI4 (7q35-q36.1) has been first investigated through targeted association analysis of *CNTNAP2*, a molecular target of *FOXP2*, i.e. the first gene implicated in human speech and language ability (see below). This analysis revealed several significant associations with three continuous language traits, namely nonword repetition, expressive and receptive language (Vernes et al., 2008). Later, novel associations of these SNPs with reading-related skills have also been reported (Newbury et al., 2011; Peter et al., 2011) and additional associations of SNPs in *CNTNAP2* have been found with further language-related traits, such as early communicative behavior (Whitehouse et al., 2011), age at first word (Alarcon et al., 2008) and age at first phrase (Anney et al., 2012). Overall, these associations suggest a wide pleiotropy of *CNTNAP2* across diverse language-related cognitive functions. *CNTNAP2* encodes an adhesion protein which is thought to have an important role in neuronal migration, dendrite outgrowth and clustering of voltage-gated ion channels at Ranvier nodes (Rodenas-Cuadrado et al., 2014).

Among SLI candidate loci, *FOXP2* gene (*Forkhead box P2;* 7q31.1) deserves a special mention. A rare missense mutation in this gene was originally discovered in a multi-generational family affected by a severe monogenic form of speech and language disorder, Childhood Apraxia of Speech (CAS) (Fisher et al., 1998; Lai et al., 2000; 2001). This disorder is characterised by difficulties in the articulation of oral speech, often accompanied by oral and written language deficits (Fisher & Scharff, 2009), and is sometimes conceived as a composite form of SLI and SSD (Pennington & Bishop, 2009). Since then, further rare variants disrupting *FOXP2* have been found to co-segregate with CAS in families, including point mutations, translocations and deletions (reviewed in Fisher & Scharff, 2009; Graham & Fisher, 2013). Mostly weak candidate SNP associations have also been reported in this gene, with continuous language traits (Rice at al., 2009) as well as with continuous reading measures and categorical RD (Peter et al. 2011; Wilcke et al. 2012), although these findings await replication. *FOXP2* encodes a transcription factor which is thought to regulate several processes within and outside of the CNS. In the CNS, it plays roles in neurite outgrowth, axon guidance, neurotransmission and synaptic plasticity (Fisher and Scharff, 2009).

Additional SLI susceptibility loci have been reported, namely *SLI2* (19q13.13-q13.41; The SLI Consortium, 2002; 2004; Monaco, 2007; Falcaro et al., 2008), *SLI3* (13q14.3-q31.1; Bartlett et al., 2002; 2004) and *SLI5* (2q36.3; Wiszniewski et al., 2013). While in *SLI2* and

SLI3 no susceptibility genes have yet been identified through genetic analyses, in *SLI5* a small heterozygous deletion in a coding sequence was found to co-segregate with a specific form of language delay in Southeast Asian families (Wiszniewski et al., 2013). This deletion -which is thought to represent a founder mutation typical of Southeast Asian populations- disrupted the *TM4SF20* gene (transmembrane 4 L six family member 20), which encodes a protein with unknown molecular functions. Therefore, further genetic analyses will be needed to confirm this gene as an SLI susceptibility locus and clarify its role in SLI etiology.

Locus	Location	Candidate genes[a]	Biological process
SLI1	16q23.1-q24	CMIP	Cytoskeletal component (potential role in Neuronal migration)
		ATP2C2	Regulation of ion levels (potential role in Synaptic plasticity, Neurotransmission, Neuronal migration)
SLI2	19q13.13-q13.41		
SLI3	13q14.3-q31.1		
SLI4	7q35-q36.1	CNTNAP2	Neuronal migration; Dendrite outgrowth; Clustering of voltage-gated ion channels
SLI5[b]	2q36.3	TM4SF20	unknown
SPCH1[b]	7q31.1	FOXP2	Neurite outgrowth; Axon guidance; Neurotransmission; Synaptic plasticity

Table 4. Loci frequently reported to be linked to SLI and language-related traits. [a] Only candidate genes implicated in SLI by genetic associations are reported. [b] These loci were not directly linked to SLI, but to related forms of language delay, and were therefore included in the present table.

Environmental influence on reading and language

Heritability data suggest that reading and language traits are also influenced by environmental factors. For RD, such variables include home language/literacy environment, socio-economic status, parental education and familial structure, as well as bioenvironmental events such as maternal health during pregnancy and lead poisoning (Grigorenko et al., 2001; Pennington & Bishop, 2009; Peterson & Pennington, 2012). By contrast, no robust evidence has been reported for environmental agents increasing susceptibility to SLI, although slight effects have been detected for some prenatal, perinatal, and neonatal factors. These include birth order (with later born having an increased SLI risk), preeclampsia (i.e. high blood pressure of the mother) during pregnancy, and in-utero exposure to high levels of testosterone

and low levels of vitamin D (Pennington & Bishop, 2009; Whitehouse et al., 2012a; 2012b; 2014). It is likely that these and other environmental factors act jointly with genetic risk factors to increase susceptibility to RD and SLI, through gene-by-environment (GxE) interactions (Pennington & Bishop, 2009). Although this field of research is still at an initial stage, it has already given promising results: possible GxE effects on RD-related traits have been reported between the candidate SNP 1259C/G in *DYX1C1* and environmental moderators, such as maternal smoke during pregnancy, birth weight and socio-economic status (Mascheretti et al., 2013).

Brain regions involved in reading and language: evidence from neuroimaging studies

In the last twenty years, an important contribution to understanding reading and language cognition and the psychopathology of RD and SLI has come from neuroimaging studies (Eicher & Gruen, 2013). Not only have these studies permitted the identification of brain regions involved in reading and language, but they also tested associations with candidate RD/SLI genes through imaging genetic analyses. These studies made use of various neuroimaging techniques (reviewed by Eicher & Gruen, 2013), aimed at testing diverse structural and functional brain measures. Such measures (presented below and, more in detail, in Chapter 6) can be considered appropriate representations of the neurobiological phenomena underlying reading and language skills, for distinct reasons. First, there is good evidence in the neuroimaging literature that structural features, e.g. gray matter volume and cortical thickness in the auditory cortex, can be directly linked to behavioral traits, e.g. auditory skills (Zatorre et al., 2012). This hypothesis has been supported also by longitudinal learning studies, where correlations have been reported between performance outcomes and brain changes (Zatorre et al., 2012). Similarly, also functional brain measures (assessed during task performance) have been linked to cognitive performance (Thompson et al., 2001; Posthuma et al., 2002). Second, independent studies indicate that both structural and functional brain measures are under significant genetic control (Thompson et al., 2001; Posthuma et al., 2002) and are highly reproducible (Thompson et al., 2010). Intriguingly, a high heritability of structural brain measures has been observed in a broad area including frontal and language-related cortical regions (Thompson et al., 2001). These data reveal a strong relationship between genes, brain structure and behavior, suggesting that highly heritable aspects of brain structure may contribute to determine individual differences in cognition.

Overall, these elements suggest that structural and functional neuroimaging measures represent appropriate endophenotypes of reading and language skills, providing an efficient mean for the investigation of RD and SLI etiology at the brain level, and increasing the power to discover genes that influence these traits (Thompson et al., 2010).

Classical neuroimaging studies

Structural Magnetic Resonance Imaging (MRI) analyses -assessing grey/white matter volume and thickness in diverse brain regions- have detected differences in the brain architecture of dyslexic and language impaired individuals, compared to non-impaired subjects (reviewed in Eicher & Gruen, 2013). These studies often found reduced gray/white matter volumes in RD/SLI cases, in brain regions such as superior temporal gyrus (STG), pars opercularis and pars triangularis in the inferior frontal gyrus (IFG) (Altarelli et al., 2014; Dole et al., 2013; Jancke et al., 2007; Badcock et al., 2012; Belton et al., 2003; Watkins et al., 2002; Brambati et al., 2006; Hoeft et al., 2007). Interestingly, these regions overlap with two brain areas that are widely considered to be implicated in language cognition, namely Broca's and Wernicke's areas (Kennison, 2013). Broca's area corresponds to pars opercularis and pars triangularis in the left IFG, while Wernicke's area overlaps with the posterior part of the left STG (Figure 2a). Structural alterations mentioned above often (but not only) affect the left hemisphere, resulting in reduced leftward asymmetries. In agreement with this, various functional MRI (fMRI) studies of RD and SLI -assessing patterns of neural activity while performing reading/language tasks- reported a reduced lateralization of both written and spoken language functions in impaired individuals, compared to controls (Eicher & Gruen, 2013; Bishop, 2013). However, whether this reduced functional asymmetry is a cause or a consequence of poor reading/language performance is still unclear (Bishop, 2013).

Other brain regions have been involved in verbal and written language skills, including cerebellum (Mariën et al., 2014), thalamus (Klostermann et al., 2013), caudate nucleus (Vargha-Khadem et al., 1998; Watkins et al., 2002; Belton et al., 2003) and many others (reviewed in Eicher & Gruen, 2013; Maisog et al., 2008). MRI studies and analyses of Fractional Anisotropy (FA) -a measure of white matter connectivity- have also highlighted the involvement of multiple fiber bundles in reading and language cognition (Vandermosten et al., 2012; Wandell & Yeatman, 2013). More specifically, anomalies in superior longitudinal, arcuate, inferior longitudinal and inferior fronto-occipital fasciculi (mainly in

the left hemisphere), as well as in the whole corpus callosum, were detected as anatomic correlates of both RD and SLI (Marino et al., 2014; Girbau-Massana et al., 2014). These fascicles are thought to be important in creating a network among brain regions involved in language capacities, as in the case of the left arcuate and inferior longitudinal fasciculi (Figure 2b), connecting IFG and STG (Eicher & Gruen, 2013; Boets et al., 2013).

Imaging genetic studies

Candidate RD/SLI genes (Table 3, 4) have been tested in different imaging genetic studies, which reported associations with some of the brain measures mentioned in the previous paragraph. The findings so far (comprehensively reviewed in Eicher & Gruen, 2013) have concerned both structural and functional phenotypes. Variants in *KIAA0319*, *DCDC2*, *DYX1C1*, *FOXP2,* and *CNTNAP2* have shown associations with gray and white matter volumes in the main language centres in the brain, including the fronto-temporal and temporo-parietal regions which largely overlap with Broca's and Wernicke's areas (Meda et al., 2008; Jamadar et al., 2011; 2013; Darki et al., 2012). In addition, increased right brain activation during reading and language tasks has been associated with variants within *KIAA0319*, *DCDC2*, *CNTNAP2* and *FOXP2*, consistent with the patterns observed in RD/SLI cases and opposed to the typical leftward pattern of activation seen in the majority of unimpaired subjects and non-carriers of these risk variants (Cope et al., 2012; Darki et al., 2012; Jamadar et al., 2011; Pinel et al., 2012; Scott-Van Zeeland et al., 2010; Whalley et al., 2011; Wilcke et al., 2012). It is worth to underline that the lines of evidence above represent oversimplifications of the findings in the field, which are actually more complex and not always perfectly consistent across studies.

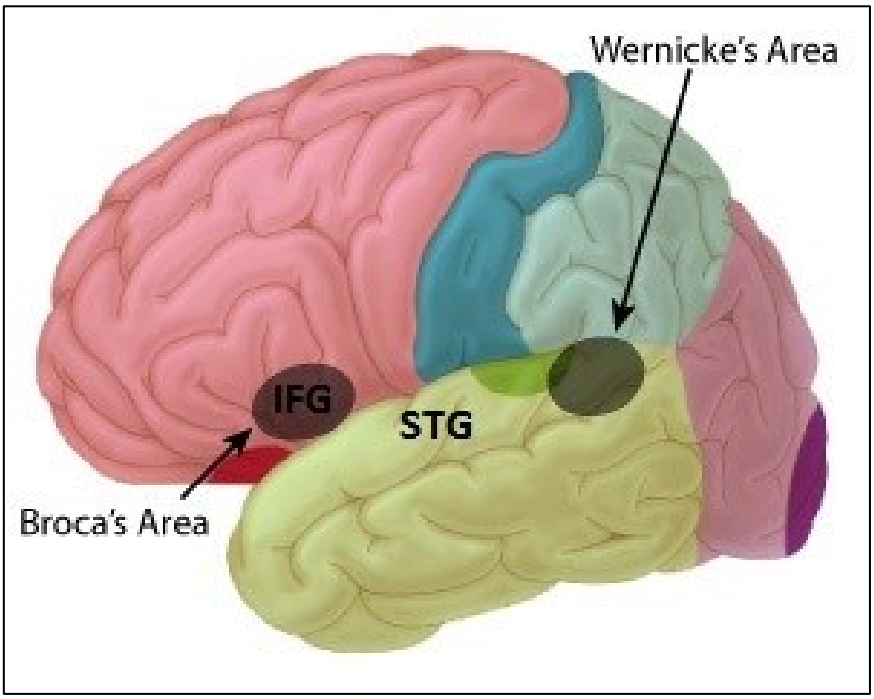
Analysis of *FOXP2* mutation carriers (coinciding with affected family members) in the CAS pedigree where this gene was originally identified (Fisher et al., 1998; Lai et al., 2000; 2001), detected functional brain anomalies associated with the causative rare mutation, including bilateral underactivation of putamen and left inferior frontal gyrus during verb generation tasks (Liégeois et al., 2003), and over-activation of the left caudate nucleus and of the ventral prefrontal region during word repetition tasks (Vargha-Khadem et al., 1998). Nonetheless, an imaging genetic analysis of various *FOXP2* SNPs -some of which had been previously associated with structural and functional neuroimaging traits in these and other brain regions- revealed no significant associations with volumetric grey and white matter measures in a

large population-based cohort, neither at the brain-wide level (in a voxel-based morphometry analysis) nor in candidate regions including caudate nucleus, cerebellum and inferior frontal cortex (investigated through structural MRI) (Hoogman et al., 2014). This suggests that the influence of *FOXP2* on brain structures may be limited to rare disruptive mutations, or that the effects of common variants in this gene are too subtle to be detected with standard volumetric techniques (Hoogman et al., 2014).

Putative autism and SLI risk variants in *CNTNAP2* were reported to be associated with more widespread and bilateral connectivity in the whole frontal cortex (Scott-Van Zeeland et al., 2010), and with reduced grey/white matter volumes in cerebellum, thalamus and right inferior longitudinal fasciculus (Tan et al., 2010). Scerri et al. (2012) also investigated the *MRPL19/GCFC2* locus, reporting an association of hypothesized dyslexia risk variants with lower verbal intelligence and with altered white matter structure in the posterior part of the corpus callosum and in the cingulum (Scerri et al., 2012).

Nonetheless, much remains to be done to fully understand the pathophysiology of reading and language at the neurobiological level. Many of the neuroimaging genetics studies in this field have been carried out on samples in the order of a few tens of subjects, which imply not only a reduced power, but also an elevated risk of false positive findings (Button et al., 2013; Hoogman et al., 2014). Therefore, further analyses in larger datasets -in the order of several thousands of subjects- are warranted to confirm these findings and to gain power to detect even very subtle genetic effects. Recently, a meta-analysis of genome-wide association studies of seven subcortical volumes was carried out in a dataset of more than 30,000 subjects, detecting five significant associations, each explaining no more than 0.52% of the phenotypic variance in the traits analysed (Hibar et al., 2015). Such small effect sizes are to be expected in future imaging genetic analyses. In addition, neuroimaging techniques will need to be refined to further increase the reliability and reproducibility of brain measures used in imaging genetic studies.

a)



b)

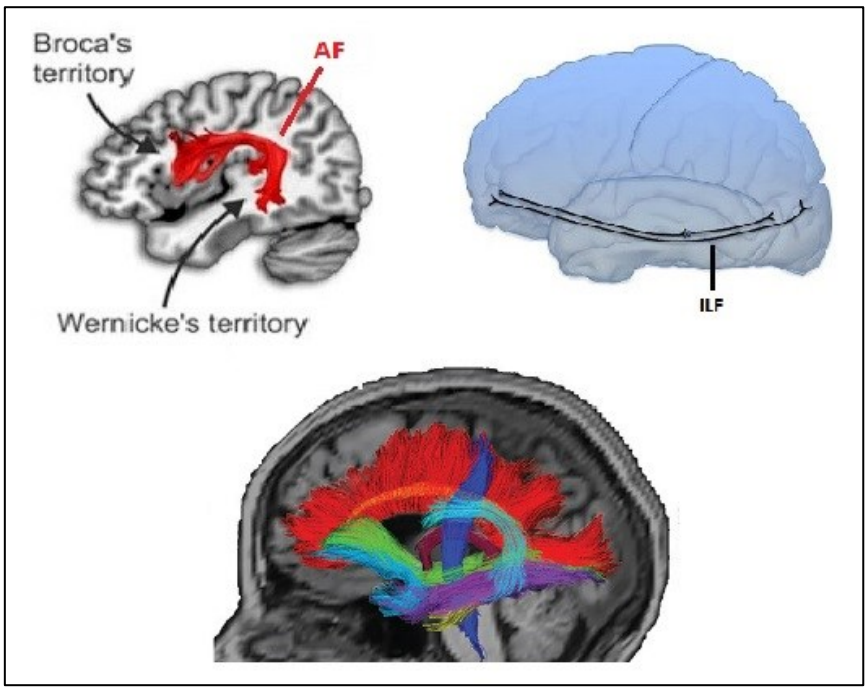


Figure 2. Location of Broca's and Wernicke's areas relative to **a)** gray and **b)** white matter architecture in the brain. The **a)** cortical areas and **b)** fiber bundles labelled in these pictures have been linked to reading and language capacities through neuroimaging evidence. Legend: IFG = inferior frontal gyrus; STG = superior temporal gyrus; AF = arcuate fasciculus; ILF = inferior longitudinal fasciculus.
Original images courtesy of **a)** http://sites.sinauer.com and **b)** López-Barroso et al. (2013); Leyden et al. (2015); Eicher & Gruen (2013).

Aims of this thesis

The main aim of this thesis was to make a contribution to understanding the overlapping genetic basis of reading and language abilities, through the investigation of genetic effects on their shared phenotypic variance.

In **Chapter 2**, I investigated the relationship between different reading and language traits, in three datasets -two from the United Kingdom and one from Colorado (US)- comprising children with reading or language problems and their siblings. Since these traits showed moderate/strong intercorrelations, I derived within each dataset a first principal component score (PC1), representing common variance in reading and language skills. Similarly, I computed a version of PC1 adjusted for performance IQ (IQ-adjusted PC1), to analyze a measure of common variance independent of nonverbal cognitive abilities. I examined the characteristics of these two principal component (PC) scores, assessing comparability across datasets, robustness and heritability, and evaluated their suitability to genetic analysis.

In **Chapter 3**, I carried out a Genome Wide Association Scan Meta-Analysis (GWASMA) of genetic variants associated with reading and language skills. This analysis, which involved the three datasets mentioned above, included Genome-Wide Association Scan with PC1/IQ-adjusted PC1 and following meta-analysis of ~5.5 million polymorphisms shared across all three datasets. I also ran a gene-based association test, in order to detect significant associations at the gene level. Finally, I assessed the patterns of pleiotropy of the two most significant association signals detected, by testing both multivariate and univariate associations of these SNPs with all the individual reading and language traits available.

In **Chapter 4**, I moved the focus onto the investigation of genes consistently implicated in RD and/or SLI, in order to detect consistency with previous findings and investigate their effects on several reading and language skills. I assessed SNP and gene-based associations of these candidate genes with PC1 and IQ-adjusted PC1, through meta-analysis of the same datasets involved in the GWASMA, and further investigated the patterns of pleiotropy of those candidate SNPs showing significant associations, as above.

In **Chapter 5**, I investigated the effect of genetic variants other than SNPs, namely Copy Number Variants (CNVs, i.e. structural variants resulting in deletion/duplication of regions larger than 1 kb in the genome), on reading and language traits. In the Colorado dataset, I first called CNVs using intensity data from DNA array (~723,000 probes) and analysed correlations between measures of CNV genomic burden and PC1/IQ-adjusted PC1, to detect

any "global" contribution of these variants to my traits of interest. Then I tested associations with PC scores through two genome-wide complementary analyses. The first analysis relied on CNV calling and later testing of association between the CNV state at each probe and PC scores, considering both deletions and duplications at each location as a single CNV state. This was aimed at detecting effects of CNVs assuming that either deletion or duplication would impact in the same way on cognitive performance. The second analysis tested association between raw intensity data for each probe and PC scores, to detect dosage-dependent effects of common multi-allelic CNVs in the genome.

In **Chapter 6**, I carried out an imaging genetic analysis of the two genes showing the strongest associations in the GWASMA, namely *FLNC* and *RBFOX2*. This was aimed at detecting potential effects of common genetic variants in these genes on brain architecture, and at assessing their compatibility with the structural brain anomalies characteristic of RD/SLI (reviewed above). In an independent Dutch population-based cohort, I analysed SNP associations with grey matter surface area and thickness of cortical regions implicated in reading and language. Both univariate and multivariate association tests were carried out with these measures, in order to detect pleiotropic genetic effects on the hypothetical "language network" formed by these regions. I also tested association with measures of asymmetry, to detect potential genetic effects on the structural lateralization of the candidate regions.

Finally, in **Chapter 7** I summarize and review the main findings of the experimental chapters (i.e. **Chapters 2** to **6**), and make a general discussion on the genetics of reading and language, with a focus on state of the art and future perspectives of this research field.

References

Adler, W.T., Platt, M.P., Mehlhorn, A.J., Haight, J.L., Currier, T.A., Etchegaray, M.A., Galaburda, A.M. & Rosen, G.D. (2013) Position of Neocortical Neurons Transfected at Different Gestational Ages with shRNA Targeted against Candidate Dyslexia Susceptibility Genes. *PLoS ONE,* **8,** e65179.

Alarcon, M., Abrahams, B., Stone, J., Duvall, J., Perederiy, J., Bomar, J., Sebat, J., Wigler, M., Martin, C., Ledbetter, D., Nelson, S., Cantor, R. & Geschwind, D. (2008) Linkage, association, and gene-expression analyses identify CNTNAP2 as an autism-susceptibility gene. *Am J Hum Genet,* **82,** 150 - 159.

Altarelli, I., Leroy, F., Monzalvo, K., Fluss, J., Billard, C., Dehaene-Lambertz, G., Galaburda, A.M. & Ramus, F. (2014) Planum temporale asymmetry in developmental dyslexia: Revisiting an old question. *Hum Brain Mapp,* **35,** 5717-5735.

American Psychiatric Association (2013) *Diagnostic and statistical manual of mental disorders (5th ed.),* Author, Washington, DC.

Andrews, W., Barber, M., Hernadez-Miranda, L.R., Xian, J., Rakic, S., Sundaresan, V., Rabbitts, T.H., Pannell, R., Rabbitts, P., Thompson, H., Erskine, L., Murakami, F. & Parnavelas, J.G. (2008) The role of Slit-Robo signaling in the generation, migration and morphological differentiation of cortical interneurons. *Dev Biol,* **313,** 648-658.

Andrews, W., Liapi, A., Plachez, C., Camurri, L., Zhang, J., Mori, S., Murakami, F., Parnavelas, J.G., Sundaresan, V. & Richards, L.J. (2006) Robo1 regulates the development of major axon tracts and interneuron migration in the forebrain. *Development,* **133,** 2243-2252.

Anney, R. & Klei, L. & Pinto, D. & Almeida, J. & Bacchelli, E. & Baird, G. & Bolshakova, N. & Bölte, S. & Bolton, P.F. & Bourgeron, T. & Brennan, S. & Brian, J. & Casey, J. & Conroy, J. & Correia, C. & Corsello, C. & Crawford, E.L. & de Jonge, M. & Delorme, R. & Duketis, E. & Duque, F. & Estes, A. & Farrar, P. & Fernandez, B.A. & Folstein, S.E. & Fombonne, E. & Gilbert, J. & Gillberg, C. & Glessner, J.T. & Green, A. & Green, J. & Guter, S.J. & Heron, E.A. & Holt, R. & Howe, J.L. & Hughes, G. & Hus, V. & Igliozzi, R. & Jacob, S. & Kenny, G.P. & Kim, C. & Kolevzon, A. & Kustanovich, V. & Lajonchere, C.M. & Lamb, J.A. & Law-Smith, M. & Leboyer, M. & Le Couteur, A. & Leventhal, B.L. & Liu, X.-Q. & Lombard, F. & Lord, C. & Lotspeich, L. & Lund, S.C. & Magalhaes, T.R. & Mantoulan, C. & McDougle, C.J. & Melhem, N.M. & Merikangas, A. & Minshew, N.J. & Mirza, G.K. & Munson, J. & Noakes, C. & Nygren, G. &--- Papanikolaou, K. & Pagnamenta, A.T. & Parrini, B. & Paton, T. & Pickles, A. & Posey, D.J. & Poustka, F. & Ragoussis, J. & Regan, R. & Roberts, W. & Roeder, K. & Roge, B. & Rutter, M.L. & Schlitt, S. & Shah, N. & Sheffield, V.C. & Soorya, L. & Sousa, I. & Stoppioni, V. & Sykes, N. & Tancredi, R. & Thompson, A.P. & Thomson, S. & Tryfon, A. & Tsiantis, J. & Van Engeland, H. & Vincent, J.B. & Volkmar, F. & Vorstman, J. & Wallace, S. & Wing, K. & Wittemeyer, K. & Wood, S. & Zurawiecki, D. & Zwaigenbaum, L. & Bailey, A.J., et al. (2012) Individual common variants exert weak effects on the risk for autism spectrum disorders. *Hum Mol Genet,* **21,** 4781-4792.

Anthoni, H., Sucheston, L., Lewis, B., Tapia-Páez, I., Fan, X., Zucchelli, M., Taipale, M., Stein, C., Hokkanen, M.-E., Castrén, E., Pennington, B., Smith, S., Olson, R., Tomblin, J.B., Schulte-Körne, G., Nöthen, M., Schumacher, J., Müller-Myhsok, B., Hoffmann, P., Gilger, J., Hynd, G., Nopola-Hemmi, J., Leppanen, P.T., Lyytinen, H., Schoumans, J., Nordenskjöld, M., Spencer, J., Stanic, D., Boon, W., Simpson, E., Mäkelä, S., Gustafsson, J.-Å., Peyrard-Janvid, M., Iyengar, S. & Kere, J. (2012) The Aromatase Gene CYP19A1: Several Genetic and Functional Lines of Evidence Supporting a Role in Reading, Speech and Language. *Behav Genet,* **42,** 509-527.

Anthoni, H., Zucchelli, M., Matsson, H., Müller-Myhsok, B., Fransson, I., Schumacher, J., Massinen, S., Onkamo, P., Warnke, A., Griesemann, H., Hoffmann, P., Nopola-Hemmi, J., Lyytinen, H., Schulte-Körne, G., Kere, J., Nöthen, M.M. & Peyrard-Janvid, M. (2007) A locus on 2p12 containing the co-regulated MRPL19 and C2ORF3 genes is associated to dyslexia. *Hum Mol Genet,* **16,** 667-677.

Badcock, N.A., Bishop, D.V.M., Hardiman, M.J., Barry, J.G. & Watkins, K.E. (2012) Co-localisation of abnormal brain structure and function in specific language impairment. *Brain Lang,* **120,** 310-320.

Baek, Seung T., Kerjan, G., Bielas, Stephanie L., Lee, Ji E., Fenstermaker, Ali G., Novarino, G. & Gleeson,

Joseph G. (2014) Off-Target Effect of doublecortin Family shRNA on Neuronal Migration Associated with Endogenous MicroRNA Dysregulation. *Neuron*, **82**, 1255-1262.

Bartlett, C.W., Flax, J.F., Logue, M.W., Smith, B.J., Vieland, V.J., Tallal, P. & Brzustowicz, L.M. (2004) Examination of Potential Overlap in Autism and Language Loci on Chromosomes 2, 7, and 13 in Two Independent Samples Ascertained for Specific Language Impairment. *Hum Hered,* **57,** 10-20.

Bartlett, C.W., Flax, J.F., Logue, M.W., Vieland, V.J., Bassett, A.S., Tallal, P. & Brzustowicz, L.M. (2002) A Major Susceptibility Locus for Specific Language Impairment Is Located on 13q21. *Am J Hum Genet,* **71,** 45-55.

Bates, T.C., Luciano, M., Castles, A., Coltheart, M., Wright, M.J. & Martin, N.G. (2007) Replication of reported linkages for dyslexia and spelling and suggestive evidence for novel regions on chromosomes 4 and 17. *Eur J Hum Genet,* **15,** 194-203.

Bates, T.C., Luciano, M., Medland, S.E., Montgomery, G.W., Wright, M.J. & Martin, N.G. (2011) Genetic variance in a component of the language acquisition device: ROBO1 polymorphisms associated with phonological buffer deficits. *Behav Genet,* **41,** 50-57.

Belton, E., Salmond, C., Watkins, K., Vargha-Khadem, F. & Gadian, D. (2003) Bilateral brain abnormalities associated with dominantly inherited verbal and orofacial dyspraxia. *Hum Brain Mapp,* **18,** 194 - 200.

Bishop, D.V. (2013) Cerebral asymmetry and language development: cause, correlate, or consequence? *Science,* **340,** 1230531.

Bishop, D.V., Laws, G., Adams, C. & Norbury, C.F. (2006) High Heritability of Speech and Language Impairments in 6-year-old Twins Demonstrated Using Parent and Teacher Report. *Behav Genet,* **36,** 173-184.

Bishop, D.V.M., Bishop, S.J., Bright, P., James, C., Delaney, T. & Tallal, P. (1999) Different Origin of Auditory and Phonological Processing Problems in Children With Language Impairment: Evidence From a Twin Study. *J Speech Lang Hear Res,* **42,** 155.

Bishop, D.V.M. & Hayiou-Thomas, M.E. (2008) Heritability of specific language impairment depends on diagnostic criteria. *Genes Brain Behav,* **7,** 365-372.

Bishop, D.V.M. & Snowling, M.J. (2004) Developmental Dyslexia and Specific Language Impairment: Same or Different? *Psychol Bull,* **130,** 858-886.

Boets, B., Op de Beeck, H.P., Vandermosten, M., Scott, S.K., Gillebert, C.R., Mantini, D., Bulthé, J., Sunaert, S., Wouters, J. & Ghesquière, P. (2013) Intact But Less Accessible Phonetic Representations in Adults with Dyslexia. *Science,* **342,** 1251-1254.

Brambati, S.M., Termine, C., Ruffino, M., Danna, M., Lanzi, G., Stella, G., Cappa, S.F. & Perani, D. (2006) Neuropsychological deficits and neural dysfunction in familial dyslexia. *Brain Res,* **1113,** 174-185.

Brambati, S.M., Termine, C., Ruffino, M., Stella, G., Fazio, F., Cappa, S.F. & Perani, D. (2004) Regional reductions of gray matter volume in familial dyslexia. *Neurology,* **63,** 742-745.

Brkanac, Z., Chapman, N., Igo, R., Jr., Matsushita, M., Nielsen, K., Berninger, V., Wijsman, E. & Raskind, W. (2008) Genome Scan of a Nonword Repetition Phenotype in Families with Dyslexia: Evidence for Multiple Loci. *Behav Genet,* **38,** 462-475.

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J. & Munafo, M.R. (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*, **14**, 365-376.

Cardon, L., Smith, S., Fulker, D., Kimberling, W., Pennington, B. & DeFries, J. (1994) Quantitative trait locus for reading disability on chromosome 6. *Science,* **266,** 276-279.

Carrion-Castillo, A., Franke, B. & Fisher, S.E. (2013) Molecular Genetics of Dyslexia: An Overview. *Dyslexia*, **19**, 214-240.

Catts, H.W., Fey, M.E., Tomblin, J.B. & Zhang, X. (2002) A Longitudinal Investigation of Reading Outcomes in Children With Language Impairments. *J Speech Lang Hear Res,* **45,** 1142-1157.

Chandrasekar, G., Vesterlund, L., Hultenby, K., Tapia-Páez, I. & Kere, J. (2013) The Zebrafish Orthologue of the Dyslexia Candidate Gene DYX1C1 Is Essential for Cilia Growth and Function. *PLoS ONE,* **8,** e63123.

Chapman, N.H., Igo, R.P., Thomson, J.B., Matsushita, M., Brkanac, Z., Holzman, T., Berninger, V.W., Wijsman, E.M. & Raskind, W.H. (2004) Linkage analyses of four regions previously implicated in

dyslexia: Confirmation of a locus on chromosome 15q. *Am J Med Genet B Neuropsychiatr Genet,* **131B,** 67-75.

Cope, N., Eicher, J.D., Meng, H., Gibson, C.J., Hager, K., Lacadie, C., Fulbright, R.K., Constable, R.T., Page, G.P. & Gruen, J.R. (2012) Variants in the DYX2 locus are associated with altered brain activation in reading-related brain regions in subjects with reading disability. *Neuroimage,* **63,** 148-156.

Cope, N., Harold, D., Hill, G., Moskvina, V., Stevenson, J., Holmans, P., Owen, M.J., O'Donovan, M.C. & Williams, J. (2005) Strong Evidence That KIAA0319 on Chromosome 6p Is a Susceptibility Gene for Developmental Dyslexia. *Am J Hum Genet,* **76,** 581-591.

Dale, P.S., Harlaar, N., Hayiou-Thomas, M.E. & Plomin, R. (2010) The Etiology of Diverse Receptive Language Skills at 12 Years. *J Speech Lang Hear Res*, **53**, 982-992.

Darki, F., Peyrard-Janvid, M., Matsson, H., Kere, J. & Klingberg, T. (2012) Three Dyslexia Susceptibility Genes, DYX1C1, DCDC2, and KIAA0319, Affect Temporo-Parietal White Matter Structure. *Biol Psychiatry,* **72,** 671-676.

de Kovel, C.G., Hol, F.A., Heister, J.G., Willemen, J.J., Sandkuijl, L.A., Franke, B. & Padberg, G.W. (2004) Genome wide scan identifies susceptibility locus for dyslexia on Xq27 in an extended Dutch family. *J Med Genet,* **41,** 652-657.

de Kovel, C.G.F., Franke, B., Hol, F.A., Lebrec, J.J.P., Maassen, B., Brunner, H., Padberg, G.W., Platko, J. & Pauls, D. (2008) Confirmation of dyslexia susceptibility loci on chromosomes 1p and 2p, but not 6p in a Dutch sib-pair collection. *Am J Med Genet B Neuropsychiatr Genet,* **147B,** 294-300.

Dole, M., Meunier, F. & Hoen, M. (2013) Gray and White Matter Distribution in Dyslexia: A VBM Study of Superior Temporal Gyrus Asymmetry. *PLoS ONE,* **8,** e76823.

Eicher, J.D. & Gruen, J.R. (2013) Imaging-genetics in dyslexia: Connecting risk genetic variants to brain neuroimaging and ultimately to reading impairments. *Mol Genet Metab*.

Fagerheim, T., Raeymaekers, P., Tonnessen, F.E., Pedersen, M., Tranebjaerg, L. & Lubs, H.A. (1999) A new gene (DYX3) for dyslexia is located on chromosome 2. *J Med Genet,* **36,** 664-669.

Falcaro, M., Pickles, A., Newbury, D.F., Addis, L., Banfield, E., Fisher, S.E., Monaco, A.P., Simkin, Z. & Conti-Ramsden, G. (2008) Genetic and phenotypic effects of phonological short-term memory and grammatical morphology in specific language impairment. *Genes Brain Behav,* **7,** 393-402.

Field, L.L., Shumansky, K., Ryan, J., Truong, D., Swiergala, E. & Kaplan, B.J. (2013) Dense-map genome scan for dyslexia supports loci at 4q13, 16p12, 17q22; suggests novel locus at 7q36. *Genes Brain Behav,* **12,** 56-69.

Fisher, S. & Scharff, C. (2009) FOXP2 as a molecular window into speech and language. *Trends Genet,* **25,** 166 - 177.

Fisher, S., Stein, J. & Monaco, A. (1999) A genome-wide search strategy for identifying quantitative trait loci involved in reading and spelling disability (developmental dyslexia). *Eur Child Adolesc Psychiatry,* **8,** S47-S51.

Fisher, S.E. & DeFries, J.C. (2002) Developmental dyslexia: genetic dissection of a complex cognitive trait. *Nat Rev Neurosci,* **3,** 767-780.

Fisher, S.E., Francks, C., Marlow, A.J., MacPhie, I.L., Newbury, D.F., Cardon, L.R., Ishikawa-Brush, Y., Richardson, A.J., Talcott, J.B., Gayan, J., Olson, R.K., Pennington, B.F., Smith, S.D., DeFries, J.C., Stein, J.F. & Monaco, A.P. (2002) Independent genome-wide scans identify a chromosome 18 quantitative-trait locus influencing dyslexia. *Nat Genet,* **30,** 86-91.

Fisher, S.E., Vargha-Khadem, F., Watkins, K.E., Monaco, A.P. & Pembrey, M.E. (1998) Localisation of a gene implicated in a severe speech and language disorder. *Nat Genet,* **18,** 168-170.

Francks, C. (2001) Mapping of genes causing dyslexia susceptibility. University of Oxford.

Francks, C., Fisher, S.E., Olson, R.K., Pennington, B.F., Smith, S.D., DeFries, J.C. & Monaco, A.P. (2002) Fine mapping of the chromosome 2p12-16 dyslexia susceptibility locus: quantitative association analysis and positional candidate genes SEMA4F and OTX1. *Psychiatr Genet,* **12,** 35-41.

Francks, C., Fisher, S.E., Marlow, A.J., MacPhie, I.L., Taylor, K.E., Richardson, A.J., Stein, J.F. & Monaco, A.P. (2003) Familial and genetic effects on motor coordination, laterality, and reading-related cognition. *Am J Psychiatry,* **160,** 1970-1977.

Francks, C., Paracchini, S., Smith, S.D., Richardson, A.J., Scerri, T.S., Cardon, L.R., Marlow, A.J., MacPhie,

I.L., Walter, J., Pennington, B.F., Fisher, S.E., Olson, R.K., DeFries, J.C., Stein, J.F. & Monaco, A.P. (2004) A 77-Kilobase Region of Chromosome 6p22.2 Is Associated with Dyslexia in Families From the United Kingdom and From the United States. *Am J Hum Genet,* **75,** 1046-1058.

Gayan, J. & Olson, R.K. (2001) Genetic and Environmental Influences on Orthographic and Phonological Skills in Children With Reading Disabilities. *Dev Neuropsychol,* **20,** 483-507.

Gayán, J. & Olson, R.K. (2003) Genetic and environmental influences on individual differences in printed word recognition. *J Exp Child Psychol,* **84,** 97-123.

Girbau-Massana, D., Garcia-Marti, G., Marti-Bonmati, L. & Schwartz, R.G. (2014) Gray–white matter and cerebrospinal fluid volume differences in children with Specific Language Impairment and/or Reading Disability. *Neuropsychologia,* **56,** 90-100.

Graham, S.A. & Fisher, S.E. (2013) Decoding the genetics of speech and language. *Curr Opin Neurobiol,* **23,** 43-51.

Grati, M.h., Chakchouk, I., Ma, Q., Bensaid, M., Desmidt, A., Turki, N., Yan, D., Baanannou, A., Mittal, R., Driss, N., Blanton, S., Farooq, A., Lu, Z., Liu, X.Z. & Masmoudi, S. (2015) A missense mutation in DCDC2 causes human recessive deafness DFNB66, likely by interfering with sensory hair cell and supporting cell cilia length regulation. *Hum Molec Genet*, 10.1093/hmg/ddv009.

Grigorenko, E.L. (2001) Developmental Dyslexia: An Update on Genes, Brains, and Environments. *J Child Psychol Psychiatry,* **42,** 91-125.

Grigorenko, E.L., Wood, F.B., Meyer, M.S., Hart, L.A., Speed, W.C., Shuster, A. & Pauls, D.L. (1997) Susceptibility loci for distinct components of developmental dyslexia on chromosomes 6 and 15. *Am J Hum Genet,* **60,** 27-39.

Hannula-Jouppi, K., Kaminen-Ahola, N., Taipale, M., Eklund, R., Nopola-Hemmi, J., Kääriäinen, H. & Kere, J. (2005) The axon guidance receptor gene ROBO1 is a candidate gene for developmental dyslexia. *PLoS Genet,* **1,** e50.

Harlaar, N., Hayiou-Thomas, M.E., Dale, P.S. & Plomin, R. (2008) Why Do Preschool Language Abilities Correlate With Later Reading? A Twin Study. *J Speech Lang Hear Res,* **51,** 688-705.

Harold, D., Paracchini, S., Scerri, T., Dennis, M., Cope, N., Hill, G., Moskvina, V., Walter, J., Richardson, A.J., Owen, M.J., Stein, J.F., Green, E.D., O'Donovan, M.C., Williams, J. & Monaco, A.P. (2006) Further evidence that the KIAA0319 gene confers susceptibility to developmental dyslexia. *Mol Psychiatry,* **11,** 1085-1091.

Hibar, D.P. & Stein, J.L. & Renteria, M.E. & Arias-Vasquez, A. & Desrivieres, S. & Jahanshad, N. & Toro, R. & Wittfeld, K. & Abramovic, L. & Andersson, M. & Aribisala, B.S. & Armstrong, N.J. & Bernard, M. & Bohlken, M.M. & Boks, M.P. & Bralten, J. & Brown, A.A. & Mallar Chakravarty, M. & Chen, Q. & Ching, C.R.K. & Cuellar-Partida, G. & den Braber, A. & Giddaluru, S. & Goldman, A.L. & Grimm, O. & Guadalupe, T. & Hass, J. & Woldehawariat, G. & Holmes, A.J. & Hoogman, M. & Janowitz, D. & Jia, T. & Kim, S. & Klein, M. & Kraemer, B. & Lee, P.H. & Olde Loohuis, L.M. & Luciano, M. & Macare, C. & Mather, K.A. & Mattheisen, M. & Milaneschi, Y. & Nho, K. & Papmeyer, M. & Ramasamy, A. & Risacher, S.L. & Roiz-Santianez, R. & Rose, E.J. & Salami, A. & Samann, P.G. & Schmaal, L. & Schork, A.J. & Shin, J. & Strike, L.T. & Teumer, A. & van Donkelaar, M.M.J. & van Eijk, K.R. & Walters, R.K. & Westlye, L.T. & Whelan, C.D. & Winkler, A.M. & Zwiers, M.P. & Alhusaini, S. & Athanasiu, L. & Ehrlich, S. & Hakobjan, M.M.H. & Hartberg, C.B. & Haukvik, U.K. & Heister, A.J.G.A.M. & Hoehn, D. & Kasperaviciute, D. & Liewald, D.C.M. & Lopez, L.M. & Makkinje, R.R.R. & Matarin, M. & Naber, M.A.M. & Reese McKay, D. & Needham, M. & Nugent, A.C. & Putz, B. & Royle, N.A. & Shen, L. & Sprooten, E. & Trabzuni, D. & van der Marel, S.S.L. & van Hulzen, K.J.E. & Walton, E. & Wolf, C. & Almasy, L. & Ames, D. & Arepalli, S. & Assareh, A.A. & Bastin, M.E. & Brodaty, H. & Bulayeva, K.B. & Carless, M.A. & Cichon, S. & Corvin, A. & Curran, J.E. & Czisch, M., et al. (2015) Common genetic variants influence human subcortical brain structures. *Nature*, 10.1038/nature14101.

Hoeft, F., Meyler, A., Hernandez, A., Juel, C., Taylor-Hill, H., Martindale, J.L., McMillon, G., Kolchugina, G., Black, J.M., Faizi, A., Deutsch, G.K., Siok, W.T., Reiss, A.L., Whitfield-Gabrieli, S. & Gabrieli, J.D.E. (2007) Functional and morphometric brain dissociation between dyslexia and reading ability. *Proc Natl Acad Sci USA,* **104,** 4234-4239.

Hoogman, M., Guadalupe, T., Zwiers, M.P., Klarenbeek, P., Francks, C. & Fisher, S.E. (2014) Assessing the effects of common variation in the FOXP2 gene on human brain structure. *Front Hum Neurosci*, **8**.

Hsiung, G.-Y.R., Kaplan, B.J., Petryshen, T.L., Lu, S. & Field, L.L. (2004) A dyslexia susceptibility locus (DYX7) linked to dopamine D4 receptor (DRD4) region on chromosome 11p15.5. *Am J Med Genet B Neuropsychiatr Genet,* **125B,** 112-119.

Huc-Chabrolle, M., Charon, C., Guilmatre, A., Vourc'h, P., Tripi, G., Barthez, M.A., Sizaret, E., Thepault, R.A., Le Gallic, S., Hager, J., Toutain, A., Raynaud, M., Andres, C., Campion, D., Laumonnier, F. & Bonnet-Brilhault, F. (2013) Xq27 FRAXA locus is a strong candidate for dyslexia: evidence from a genome-wide scan in French families. *Behav Genet,* **43,** 132-140.

Jamadar, S., Powers, N., Meda, S., Calhoun, V., Gelernter, J., Gruen, J. & Pearlson, G. (2013) Genetic influences of resting state fMRI activity in language-related brain regions in healthy controls and schizophrenia patients: a pilot study. *Brain Imaging Behav,* **7,** 15-27.

Jamadar, S., Powers, N.R., Meda, S.A., Gelernter, J., Gruen, J.R. & Pearlson, G.D. (2011) Genetic influences of cortical gray matter in language-related regions in healthy controls and schizophrenia. *Schizophr Res,* **129,** 141-148.

Jäncke, L., Siegenthaler, T., Preis, S. & Steinmetz, H. (2007) Decreased white-matter density in a left-sided fronto-temporal network in children with developmental language disorder: Evidence for anatomical anomalies in a motor-language network. *Brain Lang,* **102,** 91-98.

Kaminen, N., Hannula-Jouppi, K., Kestilä, M., Lahermo, P., Muller, K., Kaaranen, M., Myllyluoma, B., Voutilainen, A., Lyytinen, H., Nopola-Hemmi, J. & Kere, J. (2003) A genome scan for developmental dyslexia confirms linkage to chromosome 2p11 and suggests a new locus on 7q32. *Journal of medical genetics,* **40,** 340-345.

Kaplan, D.E., Gayán, J., Ahn, J., Won, T.W., Pauls, D., Olson, R.K., DeFries, J.C., Wood, F., Pennington, B.F., Page, G.P., Smith, S.D. & Gruen, J.R. (2002) Evidence for Linkage and Association with Reading Disability, on 6p21.3-22. *Am J Hum Genet,* **70,** 1287-1298.

Kennison, S.M. (2013) *Introduction to Language Development,* SAGE Publications.

Klostermann, F. (2013) Functional Roles of the Thalamus for Language Capacities. *Front Syst Neurosci,* **7**.

Lai, C., Fisher, S., Hurst, J., Levy, E., Hodgson, S., Fox, M., Jeremiah, S., Povey, S., Jamison, D., Green, E., Vargha-Khadem, F. & Monaco, A. (2000) The SPCH1 region on human 7q31: genomic characterization of the critical interval and localization of translocations associated with speech and language disorder. *Am J Hum Genet,* **67,** 357 - 368.

Lai, C., Fisher, S., Hurst, J., Vargha-Khadem, F. & Monaco, A. (2001) A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature,* **413,** 519 - 523.

Law, J., Boyle, J., Harris, F., Harkness, A. & Nye, C. (1998) Screening for primary speech and language delay: a systematic review of the literature. *Int J Lang Commun Disord,* **33 Suppl,** 21-23.

Leyden, K.M., Kucukboyaci, N.E., Puckett, O.K., Lee, D., Loi, R.Q., Paul, B. & McDonald, C.R. (2015) What does diffusion tensor imaging (DTI) tell us about cognitive networks in temporal lobe epilepsy? *Quant Imaging Med Surg*, **5**, 247-263.

Liegeois, F., Baldeweg, T., Connelly, A., Gadian, D., Mishkin, M. & Vargha-Khadem, F. (2003) Language fMRI abnormalities associated with FOXP2 gene mutation. *Nat Neurosci,* **6,** 1230 - 1237.

Logan, J., Petrill, S., Flax, J., Justice, L., Hou, L., Bassett, A., Tallal, P., Brzustowicz, L. & Bartlett, C. (2011) Genetic Covariation Underlying Reading, Language and Related Measures in a Sample Selected for Specific Language Impairment. *Behav Genet*, **41**, 651-659.

Loo, S.K., Fisher, S.E., Francks, C., Ogdie, M.N., MacPhie, I.L., Yang, M., McCracken, J.T., McGough, J.J., Nelson, S.F., Monaco, A.P. & Smalley, S.L. (2003) Genome-wide scan of reading ability in affected sibling pairs with attention-deficit//hyperactivity disorder: unique and shared genetic effects. *Mol Psychiatry,* **9,** 485-493.

López-Barroso, D., Catani, M., Ripollés, P., Dell'Acqua, F., Rodríguez-Fornells, A. & de Diego-Balaguer, R. (2013) Word learning is mediated by the left arcuate fasciculus. *Proc Natl Acad Sci USA*, **110**, 13168-13173.

Luciano, M., Lind, P.A., Duffy, D.L., Castles, A., Wright, M.J., Montgomery, G.W., Martin, N.G. & Bates, T.C. (2007) A haplotype spanning KIAA0319 and TTRAP is associated with normal variation in reading

and spelling ability. *Biol Psychiatry,* **62,** 811-817.

Maisog, J.M., Einbinder, E.R., Flowers, D.L., Turkeltaub, P.E. & Eden, G.F. (2008) A Meta-analysis of Functional Neuroimaging Studies of Dyslexia. *Ann N Y Acad Sci,* **1145,** 237-259.

Mariën, P., Ackermann, H., Adamaszek, M., Barwood, C.S., Beaton, A., Desmond, J., De Witte, E., Fawcett, A., Hertrich, I., Küper, M., Leggio, M., Marvel, C., Molinari, M., Murdoch, B., Nicolson, R., Schmahmann, J., Stoodley, C., Thürling, M., Timmann, D., Wouters, E. & Ziegler, W. (2014) Consensus Paper: Language and the Cerebellum: an Ongoing Enigma. *Cerebellum,* **13,** 386-410.

Marino, C., Meng, H., Mascheretti, S., Rusconi, M., Cope, N., Giorda, R., Molteni, M. & Gruen, J.R. (2012) DCDC2 genetic variants and susceptibility to developmental dyslexia. *Psychiatr Genet,* **22,** 25-30.

Marino, C., Scifo, P., Della Rosa, P.A., Mascheretti, S., Facoetti, A., Lorusso, M.L., Giorda, R., Consonni, M., Falini, A., Molteni, M., Gruen, J.R. & Perani, D. (2014) The DCDC2/intron 2 deletion and white matter disorganization: Focus on developmental dyslexia. *Cortex,* **57,** 227-243.

Mascheretti, S., Bureau, A., Battaglia, M., Simone, D., Quadrelli, E., Croteau, J., Cellino, M.R., Giorda, R., Beri, S., Maziade, M. & Marino, C. (2013) An assessment of gene-by-environment interactions in developmental dyslexia-related phenotypes. *Genes Brain Behav,* **12,** 47-55.

Massinen, S., Hokkanen, M.-E., Matsson, H., Tammimies, K., Tapia-Páez, I., Dahlström-Heuser, V., Kuja-Panula, J., Burghoorn, J., Jeppsson, K.E., Swoboda, P., Peyrard-Janvid, M., Toftgård, R., Castrén, E. & Kere, J. (2011) Increased Expression of the Dyslexia Candidate Gene DCDC2 Affects Length and Signaling of Primary Cilia in Neurons. *PLoS ONE,* **6,** e20580.

McArthur, G.M., Hogben, J.H., Edwards, V.T., Heath, S.M. & Mengler, E.D. (2000) On the "specifics" of specific reading disability and specific language impairment. *J Child Psychol Psychiatry,* **41,** 869–874.

McGrath, L.M., Pennington, B.F., Shanahan, M.A., Santerre-Lemmon, L.E., Barnard, H.D., Willcutt, E.G., DeFries, J.C. & Olson, R.K. (2011) A multiple deficit model of reading disability and attention-deficit/hyperactivity disorder: searching for shared cognitive deficits. *J Child Psychol Psychiatry,* **52,** 547-557.

Meda, S.A., Gelernter, J., Gruen, J.R., Calhoun, V.D., Meng, H., Cope, N.A. & Pearlson, G.D. (2008) Polymorphism of DCDC2 Reveals Differences in Cortical Morphology of Healthy Individuals—A Preliminary Voxel Based Morphometry Study. *Brain Imaging Behav,* **2,** 21-26.

Meng, H., Smith, S.D., Hager, K., Held, M., Liu, J., Olson, R.K., Pennington, B.F., DeFries, J.C., Gelernter, J., O'Reilly-Pol, T., Somlo, S., Skudlarski, P., Shaywitz, S.E., Shaywitz, B.A., Marchione, K., Wang, Y., Paramasivam, M., LoTurco, J.J., Page, G.P. & Gruen, J.R. (2005) DCDC2 is associated with reading disability and modulates neuronal development in the brain. *Proc Natl Acad Sci USA,* **102,** 17053-17058.

Monaco, A.P. (2007) Multivariate linkage analysis of specific language impairment (SLI). *Ann Hum Genet,* **71,** 660-673.

Mueller, K.L. (2012) Causation, correlation, or confound? What the comorbidity of language impairment and ADHD can tell us about the etiology of these disorders. University of Iowa.

Newbury, D.F., Bishop, D.V.M. & Monaco, A.P. (2005) Genetic influences on language impairment and phonological short-term memory. *Trends Cogn Sci*, **9**, 528-534.

Newbury, D., Fisher, S. & Monaco, A. (2010) Recent advances in the genetics of language impairment. *Genome Med*, **2**, 6.

Newbury, D., Monaco, A. & Paracchini, S. (2014) Reading and Language Disorders: The Importance of Both Quantity and Quality. *Genes (Basel),* **5,** 285-309.

Newbury, D., Winchester, L., Addis, L., Paracchini, S., Buckingham, L., Clark, A., Cohen, W., Cowie, H., Dworzynski, K., Everitt, A., Goodyer, I., Hennessy, E., Kindley, A., Miller, L., Nasir, J., O'Hare, A., Shaw, D., Simkin, Z., Simonoff, E., Slonims, V., Watson, J., Ragoussis, J., Fisher, S., Seckl, J., Helms, P., Bolton, P., Pickles, A., Conti-Ramsden, G., Baird, G., Bishop, D. & Monaco, A. (2009) CMIP and ATP2C2 modulate phonological short-term memory in language impairment. *Am J Hum Genet,* **85,** 264 - 272.

Newbury, D.F. & Monaco, A.P. (2010) Genetic advances in the study of speech and language disorders. *Neuron,* **68,** 309-320.

Newbury, D.F., Paracchini, S., Scerri, T.S., Winchester, L., Addis, L., Richardson, A.J., Walter, J., Stein, J.F.,

Talcott, J.B. & Monaco, A.P. (2011) Investigation of dyslexia and SLI risk variants in reading- and language-impaired subjects. *Behav Genet,* **41,** 90-104.

Nopola-Hemmi, J., Myllyluoma, B., Haltia, T., Taipale, M., Ollikainen, V., Ahonen, T., Voutilainen, A., Kere, J. & Widén, E. (2001) A dominant gene for developmental dyslexia on chromosome 3. *J Med Genet,* **38,** 658-664.

Nopola-Hemmi, J., Taipale, M., Haltia, T., Lehesjoki, A.-E., Voutilainen, A. & Kere, J. (2000) Two translocations of chromosome 15q associated with dyslexia. *J Med Genet,* **37,** 771-775.

Paracchini, S., Steer, C.D., Buckingham, L.L., Morris, A.P., Ring, S., Scerri, T., Stein, J., Pembrey, M.E., Ragoussis, J., Golding, J. & Monaco, A.P. (2008) Association of the KIAA0319 Dyslexia Susceptibility Gene With Reading Skills in the General Population. *Am J Psychiatry,* **165,** 1576-1584.

Paracchini, S. (2011) Dissection of genetic associations with language-related traits in population-based cohorts. *J Neurodevelop Disord,* **3,** 365-373.

Paracchini, S., Scerri, T. & Monaco, A. (2007) The genetic lexicon of dyslexia. *Annu Rev Genomics Hum Genet*, **8**, 57 - 79.

Pennington, B. & Bishop, D. (2009) Relations among speech, language and reading disorders. *Annu Rev Psych,* **60,** 283 - 306.

Pennington, B.F. (1990) The genetics of dyslexia. *J Child Psychol Psychiatry,* **31,** 193-201.

Pennington, B.F. (2006) From single to multiple deficit models of developmental disorders. *Cognition,* **101,** 385-413.

Peschansky, V.J., Burbridge, T.J., Volz, A.J., Fiondella, C., Wissner-Gross, Z., Galaburda, A.M., Turco, J.J.L. & Rosen, G.D. (2010) The Effect of Variation in Expression of the Candidate Dyslexia Susceptibility Gene Homolog Kiaa0319 on Neuronal Migration and Dendritic Morphology in the Rat. *Cereb Cortex,* **20,** 884-897.

Peter, B., Raskind, W., Matsushita, M., Lisowski, M., Vu, T., Berninger, V., Wijsman, E. & Brkanac, Z. (2011) Replication of CNTNAP2 association with nonword repetition and support for FOXP2 association with timed reading and motor activities in a dyslexia family sample. *J Neurodevelop Disord,* **3,** 39-49.

Peterson, R.L. & Pennington, B.F. (2012) Developmental dyslexia. *Lancet,* **379,** 1997-2007.

Peterson, R.L., Pennington, B.F., Shriberg, L.D. & Boada, R. (2009) What Influences Literacy Outcome in Children With Speech Sound Disorder? *J Speech Lang Hear Res,* **52,** 1175-1188.

Petryshen, T.L., Kaplan, B.J., Fu Liu, M., de French, N.S., Tobias, R., Hughes, M.L. & Leigh Field, L. (2001) Evidence for a susceptibility locus on chromosome 6q influencing phonological coding dyslexia. *Am J Med Genet,* **105,** 507-517.

Petryshen, T.L., Kaplan, B.J., Hughes, M.L., Tzenova, J. & Field, L.L. (2002) Supportive evidence for the DYX3 dyslexia susceptibility gene in Canadian families. *J Med Genet,* **39,** 125-126.

Pinel, P., Fauchereau, F., Moreno, A., Barbot, A., Lathrop, M., Zelenika, D., Le Bihan, D., Poline, J.-B., Bourgeron, T. & Dehaene, S. (2012) Genetic Variants of FOXP2 and KIAA0319/TTRAP/THEM2 Locus Are Associated with Altered Brain Activation in Distinct Language-Related Regions. *J Neurosci,* **32,** 817-825.

Platko, J.V., Wood, F.B., Pelser, I., Meyer, M., Gericke, G.S., O'Rourke, J., Birns, J., Purcell, S. & Pauls, D.L. (2008) Association of reading disability on chromosome 6p22 in the Afrikaner population. *Am J Med Genet B Neuropsychiatr Genet,* **147B,** 1278-1287.

Posthuma, D., De Geus, E.J.C., Baare, W.F.C., Pol, H.E.H., Kahn, R.S. & Boomsma, D.I. (2002) The association between brain volume and intelligence is of genetic origin. *Nat Neurosci*, **5**, 83-84.

Raitano, N.A., Pennington, B.F., Tunick, R.A., Boada, R. & Shriberg, L.D. (2004) Pre-literacy skills of subgroups of children with speech sound disorders. *J Child Psychol Psychiatry,* **45,** 821-835.

Raskind, W.H., Igo, R.P., Jr., Chapman, N.H., Berninger, V.W., Thomson, J.B., Matsushita, M., Brkanac, Z., Holzman, T., Brown, M. & Wijsman, E.M. (2005) A genome scan in multigenerational families with dyslexia: Identification of a novel locus on chromosome 2q that contributes to phonological decoding efficiency. *Mol Psychiatry,* **10,** 699-711.

Raskind, W.H., Peter, B., Richards, T.L., Eckert, M.A. & Berninger, V.W. (2013) The genetics of reading disabilities: from phenotypes to candidate genes. *Front Psychol,* **3,** 601.

Reader, R.H., Covill, L.E., Nudel, R. & Newbury, D.F. (2014) Genome-Wide Studies of Specific Language

Impairment. *Curr Behav Neurosci Rep*, **1**, 242-250.

Rice, M., Smith, S. & Gayán, J. (2009) Convergent genetic linkage and associations to language, speech and reading measures in families of probands with Specific Language Impairment. *J Neurodevelop Disord,* **1,** 264-282.

Rodenas-Cuadrado, P., Ho, J. & Vernes, S.C. (2014) Shining a light on CNTNAP2: complex functions to complex disorders. *Eur J Hum Genet,* **22,** 171-178.

Rutter, M., Caspi, A., Fergusson, D., Horwood, L.J., Goodman, R., Maughan, B., Moffitt, T.E., Meltzer, H. & Carroll, J. (2004) Sex differences in developmental reading disability - New findings from 4 epidemiological studies. *JAMA,* **291,** 2007-2012.

Scerri, T.S., Darki, F., Newbury, D.F., Whitehouse, A.J.O., Peyrard-Janvid, M., Matsson, H., Ang, Q.W., Pennell, C.E., Ring, S., Stein, J., Morris, A.P., Monaco, A.P., Kere, J., Talcott, J.B., Klingberg, T. & Paracchini, S. (2012) The Dyslexia Candidate Locus on 2p12 Is Associated with General Cognitive Ability and White Matter Structure. *PloS ONE,* **7,** e50321.

Scerri, T.S., Morris, A.P., Buckingham, L.-L., Newbury, D.F., Miller, L.L., Monaco, A.P., Bishop, D.V.M. & Paracchini, S. (2011) DCDC2, KIAA0319 and CMIP Are Associated with Reading-Related Traits. *Biol Psychiatry,* **70,** 237-245.

Schueler, M., Braun, Daniela A., Chandrasekar, G., Gee, Heon Y., Klasson, Timothy D., Halbritter, J., Bieder, A., Porath, Jonathan D., Airik, R., Zhou, W., LoTurco, Joseph J., Che, A., Otto, Edgar A., Böckenhauer, D., Sebire, Neil J., Honzik, T., Harris, Peter C., Koon, Sarah J., Gunay-Aygun, M., Saunier, S., Zerres, K., Bruechle, Nadina O., Drenth, Joost P.H., Pelletier, L., Tapia-Páez, I., Lifton, Richard P., Giles, Rachel H., Kere, J. & Hildebrandt, F. (2015) DCDC2 Mutations Cause a Renal-Hepatic Ciliopathy by Disrupting Wnt Signaling. *Am J Hum Genet*, **96**, 81-92.

Schulte-Körne, G., Grimm, T., Nöthen, M.M., Müller-Myhsok, B., Cichon, S., Vogt, I.R., Propping, P. & Remschmidt, H. (1998) Evidence for Linkage of Spelling Disability to Chromosome 15. *Am J Hum Genet,* **63,** 279-282.

Schumacher, J., Anthoni, H., Dahdouh, F., König, I.R., Hillmer, A.M., Kluck, N., Manthey, M., Plume, E., Warnke, A., Remschmidt, H., Hülsmann, J., Cichon, S., Lindgren, C.M., Propping, P., Zucchelli, M., Ziegler, A., Peyrard-Janvid, M., Schulte-Körne, G., Nöthen, M.M. & Kere, J. (2006) Strong Genetic Evidence of DCDC2 as a Susceptibility Gene for Dyslexia. *Am J Hum Genet,* **78,** 52-62.

Scott-Van Zeeland, A.A., Abrahams, B.S., Alvarez-Retuerto, A.I., Sonnenblick, L.I., Rudie, J.D., Ghahremani, D., Mumford, J.A., Poldrack, R.A., Dapretto, M., Geschwind, D.H. & Bookheimer, S.Y. (2010) Altered Functional Connectivity in Frontal Lobe Circuits Is Associated with Variation in the Autism Risk Gene CNTNAP2. *Sci Transl Med,* **2,** 56ra80.

Seshadri, S., DeStefano, A., Au, R., Massaro, J., Beiser, A., Kelly-Hayes, M., Kase, C., D'Agostino, R., DeCarli, C., Atwood, L. & Wolf, P. (2007) Genetic correlates of brain aging on MRI and cognitive test measures: a genome-wide association and linkage analysis in the Framingham study. *BMC Med Genet,* **8,** S15.

Shaywitz, S.E., Shaywitz, B.A., Fletcher, J.M.à. & Escobar, M.D. (1990) Prevalence of reading disability in boys and girls. Results of the Connecticut Longitudinal Study. *JAMA,* **264,** 998–1002.

Shriberg, L.D., Tomblin, J.B. & McSweeny, J.L. (1999) Prevalence of Speech Delay in 6-Year-Old Children and Comorbidity With Language Impairment. *J Speech Lang Hear Res,* **42,** 1461-1481.

Snowling, M., Bishop, D.V. & Stothard, S.E. (2000) Is preschool language impairment a risk factor for dyslexia in adolescence? *J Child Psychol Psychiatry,* **41,** 587–600.

Taipale, M., Kaminen, N., Nopola-Hemmi, J., Haltia, T., Myllyluoma, B., Lyytinen, H., Muller, K., Kaaranen, M., Lindsberg, P.J., Hannula-Jouppi, K. & Kere, J. (2003) A candidate gene for developmental dyslexia encodes a nuclear tetratricopeptide repeat domain protein dynamically regulated in brain. *Proc Natl Acad Sci USA,* **100,** 11553-11558.

Tallal, P. (2004) Improving language and literacy is a matter of time. *Nat Rev Neurosci,* **5,** 721-728.

Tammimies, K., Vitezic, M., Matsson, H., Le Guyader, S., Burglin, T.R., Ohman, T., Stromblad, S., Daub, C.O., Nyman, T.A., Kere, J. & Tapia-Paez, I. (2013) Molecular networks of DYX1C1 gene show connection to neuronal migration genes and cytoskeletal proteins. *Biol Psychiatry,* **73,** 583-590.

Tan, G.C.Y., Doke, T.F., Ashburner, J., Wood, N.W. & Frackowiak, R.S.J. (2010) Normal variation in fronto-

occipital circuitry and cerebellar structure with an autism-associated polymorphism of CNTNAP2. *Neuroimage,* **53,** 1030-1042.

Tarkar, A., Loges, N.T., Slagle, C.E., Francis, R., Dougherty, G.W., Tamayo, J.V., Shook, B., Cantino, M., Schwartz, D., Jahnke, C., Olbrich, H., Werner, C., Raidt, J., Pennekamp, P., Abouhamed, M., Hjeij, R., Kohler, G., Griese, M., Li, Y., Lemke, K., Klena, N., Liu, X., Gabriel, G., Tobita, K., Jaspers, M., Morgan, L.C., Shapiro, A.J., Letteboer, S.J.F., Mans, D.A., Carson, J.L., Leigh, M.W., Wolf, W.E., Chen, S., Lucas, J.S., Onoufriadis, A., Plagnol, V., Schmidts, M., Boldt, K., Uk10K, Roepman, R., Zariwala, M.A., Lo, C.W., Mitchison, H.M., Knowles, M.R., Burdine, R.D., LoTurco, J.J. & Omran, H. (2013) DYX1C1 is required for axonemal dynein assembly and ciliary motility. *Nat Genet,* **45,** 995-1003.

The SLI Consortium (2002) A genome wide scan identifies two novel loci involved in specific language impairment. *Am J Hum Genet,* **70,** 384 - 398.

The SLI Consortium (2004) Highly significant linkage to the SLI1 locus in an expanded sample of individuals affected by specific language impairment. *Am J Hum Genet,* **74,** 1225 - 1238.

Thompson, P.M., Cannon, T.D., Narr, K.L., van Erp, T., Poutanen, V.-P., Huttunen, M., Lonnqvist, J., Standertskjold-Nordenstam, C.-G., Kaprio, J., Khaledy, M., Dail, R., Zoumalan, C.I. & Toga, A.W. (2001) Genetic influences on brain structure. *Nat Neurosci*, **4**, 1253-1258.

Thompson, P.M., Martin, N.G. & Wright, M.J. (2010) Imaging genomics. *Curr Opin Neurol,* **23,** 368-373.

Tomblin, J.B., Records, N.L., Buckwalter, P., Zhang, X., Smith, E. & O'Brien, M. (1997) Prevalence of specific language impairment in kindergarten children. *J Speech Lang Hear Res,* **40,** 1245-1260.

Tran, C., Wigg, K.G., Zhang, K., Cate-Carter, T.D., Kerr, E., Field, L.L., Kaplan, B.J., Lovett, M.W. & Barr, C.L. (2014) Association of the ROBO1 gene with reading disabilities in a family-based analysis. *Genes Brain Behav,* **13,** 430-438.

Tzenova, J., Kaplan, B.J., Petryshen, T.L. & Field, L.L. (2004) Confirmation of a dyslexia susceptibility locus on chromosome 1p34-p36 in a set of 100 Canadian families. *Am J Med Genet B Neuropsychiatr Genet,* **127B,** 117-124.

Vandermosten, M., Boets, B., Wouters, J. & Ghesquière, P. (2012) A qualitative and quantitative review of diffusion tensor imaging studies in reading and dyslexia. *Neurosci Biobehav Rev,* **36,** 1532-1552.

Vargha-Khadem, F., Watkins, K.E., Price, C.J., Ashburner, J., Alcock, K.J., Connelly, A., Frackowiak, R.S.J., Friston, K.J., Pembrey, M.E., Mishkin, M., Gadian, D.G. & Passingham, R.E. (1998) Neural basis of an inherited speech and language disorder. *Proc Natl Acad Sci USA,* **95,** 12695-12700.

Velayos-Baeza, A., Toma, C., da Roza, S., Paracchini, S. & Monaco, A. (2007) Alternative splicing in the dyslexia-associated gene KIAA0319. *Mamm Genome,* **18,** 627-634.

Velayos-Baeza, A., Toma, C., Paracchini, S. & Monaco, A.P. (2008) The dyslexia-associated gene KIAA0319 encodes highly N- and O-glycosylated plasma membrane and secreted isoforms. *Hum Mol Genet,* **17,** 859-871.

Vernes, S.C., Newbury, D.F., Abrahams, B.S., Winchester, L., Nicod, J., Groszer, M., Alarcon, M., Oliver, P.L., Davies, K.E., Geschwind, D.H., Monaco, A.P. & Fisher, S.E. (2008) A functional genetic link between distinct developmental language disorders. *N Engl J Med,* **359,** 2337-2345.

Viding, E., Spinath, F.M., Price, T.S., Bishop, D.V.M., Dale, P.S. & Plomin, R. (2004) Genetic and environmental influence on language impairment in 4-year-old same-sex and opposite-sex twins. *J Child Psychol Psychiatry,* **45,** 315-325.

Wandell, B.A. & Yeatman, J.D. (2013) Biological development of reading circuits. *Curr Opin Neurobiol,* **23,** 261-268.

Wang, Y., Yin, X., Rosen, G., Gabel, L., Guadiana, S.M., Sarkisian, M.R., Galaburda, A.M. & LoTurco, J.J. (2011) Dcdc2 knockout mice display exacerbated developmental disruptions following knockdown of doublecortin. *Neuroscience,* **190,** 398-408.

Watkins, K., Dronkers, N. & Vargha-Khadem, F. (2002) Behavioural analysis of an inherited speech and language disorder: comparison with acquired aphasia. *Brain,* **125,** 452 - 464.

Whalley, H.C., O'Connell, G., Sussmann, J.E., Peel, A., Stanfield, A.C., Hayiou-Thomas, M.E., Johnstone, E.C., Lawrie, S.M., McIntosh, A.M. & Hall, J. (2011) Genetic variation in CNTNAP2 alters brain function during linguistic processing in healthy individuals. *Am J Med Genet B Neuropsychiatr Genet,* **156,** 941-

948.

Whitehouse, A.J., Bishop, D.V., Ang, Q.W., Pennell, C.E. & Fisher, S.E. (2011) CNTNAP2 variants affect early language development in the general population. *Genes Brain Behav,* **10,** 451-456.

Whitehouse, A.J., Mattes, E., Maybery, M.T., Sawyer, M.G., Jacoby, P., Keelan, J.A. & Hickey, M. (2012a) Sex-specific associations between umbilical cord blood testosterone levels and language delay in early childhood. *J Child Psychol Psychiatry,* **53,** 726-734.

Whitehouse, A.J.O., Holt, B.J., Serralha, M., Holt, P.G., Kusel, M.M.H. & Hart, P.H. (2012b) Maternal Serum Vitamin D Levels During Pregnancy and Offspring Neurocognitive Development. *Pediatrics,* **129,** 485-493.

Whitehouse, A.J.O., Shelton, W.M.R., Ing, C. & Newnham, J.P. (2014) Prenatal, Perinatal, and Neonatal Risk Factors for Specific Language Impairment: A Prospective Pregnancy Cohort Study. *J Speech Lang Hear Res,* **57,** 1418-1427.

Wilcke, A., Ligges, C., Burkhardt, J., Alexander, M., Wolf, C., Quente, E., Ahnert, P., Hoffmann, P., Becker, A., Muller-Myhsok, B., Cichon, S., Boltze, J. & Kirsten, H. (2012) Imaging genetics of FOXP2 in dyslexia. *Eur J Hum Genet,* **20,** 224-229.

Wilcke, A., Weissfuss, J., Kirsten, H., Wolfram, G., Boltze, J. & Ahnert, P. (2009) The role of gene DCDC2 in German dyslexics. *Ann Dyslexia,* **59,** 1-11.

Willcutt, E.G., Betjemann, R.S., McGrath, L.M., Chhabildas, N.A., Olson, R.K., DeFries, J.C. & Pennington, B.F. (2010) Etiology and neuropsychology of comorbidity between RD and ADHD: the case for multiple-deficit models. *Cortex,* **46,** 1345-1361.

Willcutt, E.G., Pennington, B.F., Olson, R.K., Chhabildas, N. & Hulslander, J. (2005) Neuropsychological analyses of comorbidity between reading disability and attention deficit hyperactivity disorder: in search of the common deficit. *Dev Neuropsychol,* **27,** 35-78.

Wiszniewski, W., Hunter, Jill V., Hanchard, Neil A., Willer, Jason R., Shaw, C., Tian, Q., Illner, A., Wang, X., Cheung, Sau W., Patel, A., Campbell, I.M., Gelowani, V., Hixson, P., Ester, Audrey R., Azamian, Mahshid S., Potocki, L., Zapata, G., Hernandez, Patricia P., Ramocki, Melissa B., Santos-Cortez, Regie L., Wang, G., York, Michele K., Justice, M.J., Chu, Zili D., Bader, Patricia I., Omo-Griffith, L., Madduri, Nirupama S., Scharer, G., Crawford, Heather P., Yanatatsaneejit, P., Eifert, A., Kerr, J., Bacino, Carlos A., Franklin, Adiaha I., Goin-Kochel, R.P., Simpson, G., Immken, L., Haque, Muhammad E., Stosic, M., Williams, Misti D., Morgan, Thomas M., Pruthi, S., Omary, R., Boyadjiev, Simeon A., Win, Kay K., Thida, A., Hurles, M., Hibberd, Martin L., Khor, Chiea C., Van Vinh Chau, N., Gallagher, Thomas E., Mutirangura, A., Stankiewicz, P., Beaudet, Arthur L., Maletic-Savatic, M., Rosenfeld, Jill A., Shaffer, Lisa G., Davis, Erica E., Belmont, John W., Dunstan, S., Simmons, C.P., Bonnen, P.E., Leal, Suzanne M., Katsanis, N., Lupski, James R. & Lalani, Seema R. (2013) TM4SF20 Ancestral Deletion and Susceptibility to a Pediatric Disorder of Early Language Delay and Cerebral White Matter Hyperintensities. *Am J Hum Genet,* **93,** 197-210.

Chapter 2: Phenotypic analysis of reading and language traits

Part of this chapter has been published in

Gialluisi, A., Newbury, D.F., Wilcutt, E.G., Olson, R.K., DeFries, J.C., Brandler, W.M., Pennington, B.F., Smith, S.D., Scerri, T.S., Simpson, N.H., The SLI Consortium, Luciano, M., Evans, D.M., Bates, T.C., Stein, J.F., Talcott, J.B., Monaco, A.P., Paracchini, S., Francks, C. & Fisher, S.E. (2014) Genome-wide screening for DNA variants associated with reading and language traits. *Genes, Brain and Behavior*, 13(7), 686-701.

Abstract

Reading and language skills are complex cognitive traits showing shared genetic and environmental influences. However, so far genetic studies have mainly focused on their corresponding deficits, namely Reading Disability (RD) and Specific Language Impairment (SLI), and on individual reading- and language-related traits. Here, we investigated the relationship between several continuous reading and language traits through an exploratory phenotype analysis, in three richly characterized datasets of individuals with histories of reading or language problems and their siblings. Within each dataset we observed moderate to high cross-traits correlations, hence we derived a first principal component score (PC1) representing common variance in reading and language skills. We describe the characteristics of PC1 within each dataset, including loadings on the specific traits, correlations with IQs and maximum heritability estimates based on sibling correlations.

PC1 showed a broad pattern of loadings across the traits and explained a substantial proportion of their common variance (52-75%) in all the datasets. Dropping one or more traits from our PC1 computation did not substantially affect the resulting PC1 scores. Furthermore, moderate correlations with nonverbal cognitive skills were reported ($r = 0.21$-0.46), which prompted us to compute also a PC1 score adjusted for performance IQ (IQ-adjusted PC1). Finally, PC1 showed moderate to high heritabilities in all the datasets (0.29-0.84), in line with previous heritability estimates on RD, SLI and continuous reading and language traits.

In conclusion, our results support the existence of a common phenotypic variance in reading and language skills, which is partly shared with general cognitive abilities and moderately influenced by genetic factors. We demonstrated that PC1 is an appropriate proxy measure of this common variance, characterized by robustness, heritability and broad comparability across phenotypically heterogeneous datasets. These elements make PC1 a suitable trait for genetic analysis aimed at detecting variants with pleiotropic effects on reading and language traits.

Introduction

Reading and language skills are complex cognitive traits (i.e. influenced by several genetic and environmental factors) that are strictly intertwined and show strong intercorrelations, attributed to common genetic and environmental influences (Harlaar et al., 2008; Logan et al., 2011). Accordingly, Reading Disability (RD, or Developmental Dyslexia) and Specific Language Impairment (SLI) are often comorbid (Snowling et al. 2000; McArthur et al. 2000). These conditions show comorbidities also with other neurodevelopmental disorders, such as Attention Deficit Hyperactivity Disorder (ADHD) (Wilcutt et al. 2010; Pennington, 2006; Mueller, 2012). These comorbidities (reviewed in Chapter 1) suggest the presence of shared neurobiological bases for reading and language skills. As an example, some neuropsychological theories emphasize the role of phonological processing deficits in the etiology of both RD and SLI (see Chapter 1). Nonetheless, multiple deficits are thought to contribute to these disorders, including language syntax and phonological short term memory deficits for SLI, and orthographic coding and rapid automatic naming deficits for RD (see Bishop and Snowling, 2004; Pennington & Bishop, 2009 for a review). Some of these deficits are also shared with other neurodevelopmental disorders, as in the case of rapid automatic naming and central processing speed, which are impaired both in RD and in ADHD (Willcutt et al., 2005; Willcutt et al., 2010; McGrath et al., 2011). These lines of evidence have led to hypothesize a complex network of interconnections among reading, language and other neurodevelopmental cognitive domains (e.g. attention).

Cognitive traits underlying reading and language are generally characterized by moderate to high heritability (*h*, i.e. the proportion of phenotypic variance explained by additive genetic factors), as reported by several independent studies. Francks et al. (2003) estimated heritabilities of different reading- and language-related traits, namely word reading, spelling, orthographic coding, phonological decoding and phoneme awareness, assessing full sibling correlation (i.e. familiality) in 265 nuclear sibling-pair families from the UK. In this sample, which is part of the UK-RD dataset analysed in the present chapter (see *Subjects and Methods* section), each family contained at least one reading-disabled proband and heritabilities were computed as twice the familiality coefficients. Reading and language traits exhibited heritabilities of 0.39-0.66 (Francks et al., 2003). These estimations include also the proportion of phenotypic variance attributable to shared environmental factors and hence represent an upper limit of the real heritability values. Therefore, they will be indicated as "maximum heritability" estimates hereafter. Comparable heritability values were computed in

a similar cohort, made up of 515 twins recruited in Colorado for a school history of RD and their unimpaired cotwins (DeFries et al., 1997). In this dataset, which is part of the CLDRC dataset analysed in the present chapter (see *Subjects and Methods* section), Gayán & Olson (2001) reported heritabilities of 0.46–0.72 for phoneme awareness, 0.57–0.59 for word recognition, 0.60–0.71 for phonological decoding, and 0.55–0.67 for orthographic coding measures. These estimates were based on the DeFries-Fulker regression method (DeFries & Fulker 1985), which exploits twin data to evaluate the heritability of extreme deficits for a continuous trait of interest. This is also defined as "group heritability" (h_g), to distinguish it from estimates of heritability based on interindividual variation in the normal range of ability. In the same study, an assessment of pairwise bivariate heritabilities among word recognition, orthographic coding, phonological decoding and phoneme awareness yielded significant estimates, providing evidence for common genetic etiologies of deficits across these reading and language skills (Gayán & Olson, 2001). Further analyses on a similar sample (on 440 pairs of twins) aimed at estimating heritability of latent reading and language traits (through ACE Cholesky Decomposition Models; Gayán & Olson, 2003). These produced slightly higher h estimates -0.83 for phoneme awareness, 0.85 for word recognition, 0.8 for phonological decoding and 0.87 for orthographic coding- which were hypothesized to be due to non-additive genetic effects (Gayán & Olson, 2003). These results are comparable with the assessments of heritability for composite/component scores derived from several reading and language traits. For a composite score of overall reading performance (based on word recognition, spelling and comprehension and described in detail in Chapter 5), heritability was estimated to be approximately 0.5-0.6 in the CLDRC dataset (Friend et al., 2010; DeFries & Gillis, 1993). In a comparative analysis of the Colorado and UK samples mentioned above, Francks (2001) estimated heritability for the first principal component (PC) scores derived from several reading and language traits. The first PC in the UK dataset, derived from word reading, spelling, orthographic coding, phoneme awareness and phonological decoding, showed a maximum heritability estimate of 0.68 (estimated through variance component approach in SOLAR; Almasy & Blangero, 1998). Comparably, the first PC in the Colorado dataset, computed from a similar set of measures including also rapid automatic naming, showed a maximum heritability of 0.54 (Francks, 2001).

For SLI and relevant language traits, even higher heritabilities have been reported by independent works (Bishop et al., 1995; 1999; Tomblin and Buckwalter, 1998; Viding et al., 2004; see Bishop & Hayiou-Thomas, 2008 for a review). In these studies, mainly group

heritabilities were computed, by comparing concordance of the disease status between pairs of monozygotic (MZ) and pairs of dizygotic (DZ) twins (higher concordance rates in MZ twins suggest the presence of a genetic etiology for the disorder of interest). These estimates, generally ranging between ~0.4 and ~0.75, were confirmed for a nonword repetition measure in a sample of British children affected by SLI and their cotwins, and in a sample of twinships from the British general population (Bishop et al., 1999). Common heritability for this trait was $h = 0.71$; while group heritability was $h_g > 1$, suggesting the presence of non-additive genetic effects on SLI. Nonetheless, in the same study, an auditory processing score showed non-significant heritability, both in the general population and in the SLI sample ($h_g \sim 0.22$; Bishop et al., 1999). In a large study on 4,892 12-year-old twin pairs, four measures of receptive language development -including vocabulary, listening grammar, figurative language, and making inferences- showed moderate genetic influence (heritability 0.25-0.36; Dale et al., 2010). Higher estimates were reported for a latent factor score for language, based on the common variance among these measures ($h = 0.59$ in ACE model; Dale et al., 2010).

To summarize, heritability data published so far support the view that a substantial proportion of phenotypic variance in reading and language is attributable to genetic factors, and make reading and language traits appropriate for genetic studies aimed at gaining more knowledge on the genetic underpinnings of RD and SLI. In spite of evidence pointing at shared cognitive deficits for reading and language disabilities (see above and Chapter 1), genetic studies in the past have mainly investigated single deficits (either RD or SLI) and individual reading/language traits (see Carillon-Castillo et al., 2013; Newbury & Monaco, 2010 for a review).

Only recently two Genome Wide Association Scans (GWAS) were run with the aim of identifying pleiotropic variants with an effect on both reading and language skills. In a GWAS meta-analysis (GWASMA) on quantitative reading and language traits in two population based cohorts (N~6,500), three measures were analyzed: word reading, nonword repetition and a proxy measure of reading-spelling ability (Luciano et al., 2013). The latter trait consisted of a principal component score derived from regular-word reading, irregular-word reading, nonword reading and spelling in one cohort, which was meta-analyzed with a composite measure of word reading, nonword reading and spelling in the other cohort. All the measures were residualized against sex, age and performance IQ (see Luciano et al., 2013 for further details). More recently, a case-control GWAS compared a reduced number of comorbid RD-SLI cases (N=174) to general population controls (N~4,100; Eicher et al.,

2013). In this study, RD cases were defined as scoring at least 1 standard deviation (SD) below the mean of the general population for at least three out of five tasks, including phoneme deletion at age 7, single word reading at age 7 and 9, nonword reading at age 9 and reading comprehension at age 9. SLI cases were instead defined as subjects scoring at least 1 SD below the mean for at least two out of three language traits, namely phoneme deletion at age 7, verbal comprehension at age 8 and nonword repetition at age 8. All the subjects involved in the study presented with full scale IQ $\geq$ 75 (Eicher et al., 2013). The different phenotypes analysed in these works reflected two complementary GWA strategies, one aimed at identifying pleiotropic variants affecting variance in reading and language skills across a broad range of variation, and the other aimed at detecting variants associated with poor performance, focused on the lower tail of the distribution in reading and language traits (the results will be discussed in detail in Chapter 3). Similarly to these recent GWAS studies of reading/language performance (Luciano et al. 2013; Eicher et al., 2013), we aimed to carry out a GWAS meta-analysis for genetic variants influencing reading and language abilities (described in Chapter 3), including three long-established datasets which comprised children with reading or language problems, along with their siblings. This approach complements the above mentioned GWAS studies as it investigates continuous trait variance across a broad range of reading and language abilities, with an enrichment for the poor performing tail of the distribution. At the same time, it does not apply any arbitrary dichotomy between RD/SLI cases and controls, which often constitutes a source of heterogeneity across studies (Pennington & Bishop, 2009; Raskind et al., 2013).

In the present chapter, we investigated the relationship between the reading and language traits available for our GWAS meta-analysis. We observed generally strong intercorrelations across these traits, which suggested the existence of a notable proportion of phenotypic variance shared between reading and language abilities. Hence we derived a first principal component score, representing common variance in reading and language skills, within each dataset. We examined the characteristics of this score, assessing comparability across datasets, robustness and heritability, and concluded that it was an appropriate measure of this common variance and a trait suitable to genetic analysis.

A notable part of the variance in reading and language abilities is shared with general cognition (Gayán & Olson, 2003), while another part of this variance is independent of IQ (Pennington & Bishop, 2009). As in our study we identified moderate correlations with general cognitive abilities, we also computed an IQ-adjusted version of the first principal

component score and assessed its characteristics in this chapter, to analyse the common variance in reading and language traits independent of general intelligence.

Subjects and Methods

Datasets

UK-RD

This dataset comprised children diagnosed with RD, and their siblings, collected at the Dyslexia Research Centre clinics in Oxford and Reading, or the Aston Dyslexia and Development Clinic in Birmingham, United Kingdom. Ethical approval was acquired from the Oxfordshire Psychiatric Research Ethics Committee (OPREC O01.02) and written informed consent of the participants (or their parents) was obtained. The total number of participants was 983, mean age 11.7 years, age range 5-31, from 608 independent nuclear families. All children, regardless of diagnosis, were administered psychometric tests of reading- and language-related abilities, as well as assessments of verbal and non-verbal IQ (details further below).

SLIC

The SLI Consortium dataset comprised children affected by SLI, along with their siblings, recruited from five centres across the UK; The Newcomen Centre at Guy's Hospital, London (now called Evelina Children's Hospital); the Cambridge Language and Speech Project (CLASP); the Child Life and Health Department at the University of Edinburgh; the Department of Child Health at the University of Aberdeen; and the Manchester Language Study, as described in previous reports by the SLI Consortium (SLIC 2002; 2004; Falcaro et al., 2008; Newbury et al., 2009). This sample included 49 families from the Guy's Hospital, London cohort which had not been included in previous SLI Consortium studies. Ethical agreement was given by local ethics committees of the hospitals involved in the consortium, and all subjects provided informed consent. All children in this sample were assessed for a number of reading- and language-related traits (see below) regardless of their language ability. For this study we obtained data for affected probands and their available siblings, for a total of 548 participants, mean age 10 years, age range 5-19, from 288 independent nuclear families.

CLDRC

The Colorado Learning Disabilities Research Centre (CLDRC) dataset was derived from an ongoing study on the etiology of learning disabilities run in 27 school districts in Colorado, USA (DeFries et al., 1997; Willcutt et al., 2005). Pairs of twins were initially recruited based on a school report of RD, ADHD or other learning disabilities in one or both of the twins; they were then administered a number of psychometric tests for several learning-related skills, along with their additional co-siblings, and DNA was collected for genetic studies. The Institutional Review Boards of the University of Nebraska Medical Center and of the University of Colorado at Boulder had approved the protocol, and written informed consent of the participants (or their parents) was obtained.

For the present study, for MZ twin pairs, we selected one child per pair based on the maximum availability of reading- and language-related trait data, or otherwise randomly. The sample of twins and siblings available for this study comprised 749 participants in total, mean age 11.7 years, age range 8-19, from 343 unrelated twinships/sibships. Of these, 266 of the twinships/sibships (a total of 585 participants) were originally recruited via a proband with a history of RD, and 77 of the twinships/sibships (164 participants in total) were originally recruited via a proband with a history of ADHD. The two subsets are indicated hereafter as CLDRC-RD and CLDRC-ADHD.

Reading and language measures

Table 1 lists the reading- and language-related traits that were assessed in the different datasets, as detailed in prior publications (Compton et al., 2001; Friend & Olson, 2010; Francks et al., 2004; SLIC 2002; 2004). Further information on these measures is given in Tables S1a, b, c. To remove outliers, trait scores were excluded when they were more than 3 standard deviations from the relevant sample mean. Subjects with three or more such outliers were excluded from the dataset (one participant in UK-RD and one in CLDRC-RD). Reading/language traits had been previously age-adjusted according to normative data (Compton et al., 2001; Friend & Olson, 2010; Francks et al., 2004; SLIC 2002; 2004). When a measure differed significantly from normality we performed a within-dataset rank-normalization to attain normality and improve the suitability for principal components analysis (see Table S1a, b, c for details). We also excluded subjects showing full scale IQ < 70 (one participant from CLDRC-RD, and four participants from SLIC). This left 564

subjects in CLDRC-RD, 958 in UK-RD, 498 in SLIC and 163 in CLDRC-ADHD, which were used for the computation of the First Principal Component. To correct for relatedness of subjects, pairwise trait correlations within each dataset were calculated as the median Pearson's r correlation over 100 repeat random samplings of one individual from each independent sibship, using R (R core Team, 2013, http://www.r-project.org/). Similarly, we calculated correlations of the reading and language traits available with the IQ measures (both verbal and performance IQ).

First Principal Component scores

Computation

In light of the moderate/high cross-traits correlations detected (see *Results* section below), we derived the First Principal Component (PC1) from all of the language- and reading-related traits available in each dataset, through the SPSS® 20.0 Factor Analysis (Principal Component extraction method, hereafter called PCA).

This reduced our correlated measures into a smaller set of latent variables (factors or principal components) that can explain the maximum amount of shared variance (Field, 2005). In each dataset, only linear components with Eigenvalue>1 were extracted, allowing for correlation among the components (oblique rotation, *direct oblim* method) and excluding subjects with any missing measure (*missing listwise* option). A Kaiser-Meyer-Olkin measure of sampling adequacy and a Bartlett's test of sphericity were run in all the PCAs. These tests revealed a high common variance (KMO = 0.8-0.9) and a significant interdependence (Bartlett's test p-value < 0.05) among the variables examined in each dataset, further justifying the PCAs. We also derived a first principal component score within each dataset from only word reading and spelling, because these were the only measures available in all datasets and therefore provided a possibility to match traits as closely as possible across datasets. The first PC derived from word reading and spelling is referred to as $PC1_{read}$ hereafter.

Trait[a]	Description (ability assessed)	CLDRC-RD (564)	UK-RD (958)	SLIC (498)	CLDRC-ADHD (163)
WRead	Reading real words	x (0.918)	x (0.918)	x (0.902)	x (0.871)
WSpell	Spelling real words	x (0.813)	x (0.852)	x (0.862)	x (0.764)
PD	Ability to convert letter strings into sounds, according to given phonetic rules	x (0.895, 0.861)[b]	x (0.809)		x (0.821, 0.729)[b]
PA	Ability to recognize and manipulate speech sounds (phonemes)	x (0.801)	x[c]		x (0.744)
OC	Ability to recognize a word as an orthographic unit and to retrieve the corresponding phonological form	x (0.764)	x (0.888)		x (0.644)
NWR	Ability to repeat nonsense words orally presented	x (0.493)		x (0.665)	x (0.355)
ELS	Sentence recalling and production (expressive domain of language)			x (0.856)	
RLS	Listening and auditory comprehension (receptive domain of language)			x (0.837)	
VIQ	Verbal reasoning	x	x	x	x
PIQ	Logical reasoning	x	x	x	x
PC1 (N)	Common variance in reading and language skills	544	914	245	159
IQ-adjusted PC1 (N)	Common variance in reading and language skills, not shared with general cognitive abilities	544	878	245	159

Table 1. Phenotypic traits available (when labeled by "x") and measures used for PC1 extraction within each dataset (labeled with relative loadings on PC1 in parentheses). Sample sizes of the datasets before Principal Component Analysis are reported in the header row. Numbers of subjects with PC1/IQ-adjusted PC1 measures available within each dataset are reported in the bottom rows (since we excluded participants with at least one missing measure among the traits involved in PCA).
[a] Legend: WRead = word reading; WSpell = word spelling; PD = phonological decoding; PA = phoneme awareness; OC = orthographic coding; NWR = nonword repetition; ELS/RLS = expressive/receptive language score; VIQ/PIQ = verbal/performance IQ. [b] Loadings of nonword reading and phonological choice (respectively) on PC1s. [c] Trait excluded from the PCA due to the low number of measures available.

Robustness and composition of PC scores

We further assessed the robustness of our PC1 scores by evaluating the impact of adding/removing one or more traits in the PCAs. First we added/removed one or two measures before the PCA in a dataset and computed the resulting PC1 score (hereafter called "experimental" PC1).We did so for those measures with a high missing rate in the datasets, which would have therefore entailed a non-negligible reduction in the total sample size of the GWASMA. Then we assessed the correlation between the "original" PC1 score (extracted from the reading and language traits available in the datasets, as described in Table 1) and the "experimental" PC1 score. This correlation was computed as the median Pearson's r correlation between the two PC1 scores, over 100 repeat random samplings of one individual from each independent sibship. In UK-RD, the original PC1 -derived from word reading, spelling, phonological decoding and orthographic coding- was compared to an "experimental" PC1 score computed also from phoneme awareness. Similarly, the original PC1 in SLIC -derived from word reading, spelling, expressive and receptive language and nonword repetition- was compared with two experimental PC1 scores, one based only on word reading, spelling and nonword repetition, and the other one based on language scores and nonword repetition. In the CLDRC datasets, there were low missing rates for all the measures (see Table S2), therefore we decided to use all the reading and language traits available in PC1 computation and no correlations with PC1 scores after adding/removing one or more traits were assessed. Further details on the composition of PC1 scores within each dataset are reported in Table 1 and in *Supplementary Material S2*.

Correlation patterns and IQ-adjustment of PC scores

Correlations of PC1 and $PC1_{read}$ with IQ measures were calculated, within each dataset, as the median Pearson's r correlation over 100 repeat random samplings of one individual from each independent sibship (as for all pairwise trait correlations; see above).

To remove the variance shared between general (nonverbal) cognitive abilities and measures of reading and language, we obtained residuals from regressing PC1 and $PC1_{read}$ against performance IQ, again separately within each dataset. The resulting residual scores will be called IQ-adjusted PC1 and IQ-adjusted $PC1_{read}$ hereafter.

Pairwise correlations among all PC scores (both before and after IQ-adjustment) were also computed within each dataset, as above.

Analysis of familiality/heritability of PC scores

We assessed familiality and heritability of principal component scores -namely PC1, $PC1_{read}$, IQ-adjusted PC1 and IQ-adjusted $PC1_{read}$- by calculating full sibling correlations, separately within datasets. Since some sibships contained more than two individuals and others contained a single subject after QC, we first computed Pearson's r sibling correlations over 100 random samplings of one sibling pair from each family with two or more siblings. Then we computed median values within each dataset, and multiplied these familiality parameters by two to have estimates of common heritability of our PC scores. Although this kind of estimate may constitute an overestimation of heritability, as it also includes the fraction of phenotypic variance explained by shared environmental factors, it was still useful to evaluate the extent to which genetic factors affect our traits of interest.

Results

Correlation patterns of single reading/language traits

Moderate to high cross-phenotypic correlations were detected for most of the reading and language traits in the datasets (Tables 2a, b, c, d), justifying the extraction of principal component scores (PC1 and $PC1_{read}$, see below). Correlation coefficients ranged from 0.22 to 0.85 in CLDRC-RD, from 0.53 to 0.77 in UK-RD, from 0.41 to 0.87 in SLIC, and from 0.09 to 0.78 in CLDRC-ADHD.

2a)

Trait[a]	WRead	WSpell	PD (NWRead)	PD (PC)	PA	OC	NWR
WRead	1	0.768	0.851	0.736	0.652	0.661	0.396
WSpell		1	0.671	0.583	0.503	0.691	0.286
PD (NWRead)			1	0.759	0.694	0.596	0.328
PD (PC)				1	0.73	0.606	0.336
PA					1	0.445	0.422
OC						1	0.219
NWR							1

2b)

Trait[a]	WRead	WSpell	PD	PA	OC
WRead	1	0.749	0.634	0.552	0.77
WSpell		1	0.53	0.537	0.662
PD			1	0.656	0.595
PA				1	0.625
OC					1

2c)

Trait[a]	WRead	WSpell	NWR	ELS	RLS
WRead	1	0.869	0.463	0.679	0.647
WSpell		1	0.412	0.618	0.584
NWR			1	0.536	0.433
ELS				1	0.769
RLS					1

2d)

Trait[a]	WRead	WSpell	PD (NWRead)	PD (PC)	PA	OC	NWR
WRead	1	0.622	0.779	0.542	0.566	0.454	0.279
WSpell		1	0.491	0.405	0.492	0.572	0.157
PD (NWRead)			1	0.556	0.629	0.373	0.171
PD (PC)				1	0.499	0.419	0.091
PA					1	0.291	0.219
OC						1	0.089
NWR							1

Table 2. Pairwise trait correlations of reading and language measures in **a)** CLDRC-RD, **b)** UK-RD, **c)** SLIC, and **d)** CLDRC-ADHD datasets. These were computed separately within each dataset, as the median Pearson's correlation over 100 repeat random samplings of one individual from each independent sibship.
[a] Legend: WRead = word reading; WSpell = word spelling; PD = phonological decoding (NWRead = nonword reading and PC = phonological choice); PA = phoneme awareness; OC = orthographic coding; NWR = nonword repetition; ELS/RLS = expressive/receptive language score.

First Principal Component scores

The proportion of total phenotypic variance explained by PC1 was 75.3% in UK-RD, 68.6% in SLIC, 64.5% in CLDRC-RD, and 52.0% in CLDRC-ADHD. In all the datasets PC2 explained no more than 13% of the total variance. All of the PC1s showed a broad pattern of loadings across the traits (Table 1). Furthermore, dropping one or more traits from our PC1 computation did not substantially affect the resulting PC1 scores. Correlation in UK-RD

between the original PC1 -derived from word reading, spelling, phonological decoding and orthographic coding- and an experimental PC1 based also on phoneme awareness, was high (Pearson's r = 0.99). Similarly, strong correlations were observed in SLIC between the original PC1 -computed from word reading, spelling, expressive and receptive language scores and nonword repetition- and two experimental PC1 scores, one computed after removal of language scores and one computed after removal of word reading and spelling (Pearson's r = 0.95 and 0.94, respectively). The total number of participants for which we finally obtained PC1 data (i.e. all datasets combined) was 1,862.

The proportion of variance in word reading and spelling explained by $PC1_{read}$ was 86.9% in UK-RD, 88% in CLDRC-RD, 93.4% in SLIC and 80.1% in CLDRC-ADHD. As only two measures were used to construct $PC1_{read}$, then these measures loaded equally onto this component, and the loadings were high in all datasets (≥ 0.9, Table 3). The total number of subjects across all datasets for $PC1_{read}$ was 1,913.

Trait[a]	CLDRC-RD	UK-RD	SLIC	CLDRC-ADHD
WRead	0.938	0.932	0.967	0.895
WSpell	0.938	0.932	0.967	0.895
$PC1_{read}$ (N)	558	925	271	159
IQ-adjusted $PC1_{read}$ (N)	558	888	270	159

Table 3. Phenotypic measures used for $PC1_{read}$ computation within each dataset and relative loadings on $PC1_{read}$. Numbers of subjects with $PC1_{read}$/IQ-adjusted $PC1_{read}$ measures available within each dataset are reported in the bottom rows.
[a] Legend: WRead = word reading; WSpell = word spelling.

Correlation patterns with IQ and IQ-adjustment of PC scores

We observed moderate to high correlations with IQs, both for individual reading/language traits and for principal component scores (Tables 4a, b, c, d). As expected, verbal IQ generally showed higher correlations with the reading- and language-related traits than performance IQ, although the latter measure was also moderately correlated, more prominently in SLIC (Table 4c). Among the principal component scores, PC1 showed correlations with performance IQ of 0.32 in both CLDRC-RD and UK-RD, 0.46 in SLIC and 0.21 in CLDRC-ADHD, while $PC1_{read}$ showed correlations of 0.32 in CLDRC-RD, 0.33 in UK-RD, 0.39 in SLIC and 0.25 in CLDRC-ADHD. Similarly to individual reading/language traits, correlations of PC scores with verbal IQ were higher than those with performance IQ: PC1 showed correlations of 0.53 in CLDRC-RD, 0.43 in UK-RD, 0.79 in SLIC, and 0.58 in

CLDRC-ADHD, while for $PC1_{read}$ correlations were 0.56 in CLDRC-RD, 0.42 in UK-RD, 0.68 in SLIC, and 0.57 in CLDRC-ADHD.

In the consequent IQ-adjustment of PC scores, a measure of performance IQ was not available for 36 of the 1,862 participants with PC1 score available (in all the datasets combined), and therefore the total sample size for IQ-adjusted PC1 analysis was 1,826. Similarly, the total sample size for IQ-adjusted $PC1_{read}$ analysis was reduced to 1,875 (from 1,913 $PC1_{read}$ measures available in all the datasets combined).

Cross-trait correlations of PC scores

The correlations between PC1 and $PC1_{read}$ were high in all datasets, both before (Pearson's r = 0.92 in CLDRC-RD, 0.95 in UK-RD, 0.91 in SLIC and 0.92 in CLDRC-ADHD) and after IQ-adjustment (Pearson's r = 0.92 in CLDRC-RD, 0.94 in UK-RD, 0.89 in SLIC and 0.91 in CLDRC-ADHD), as reported in Tables 5a, b, c, d. These tables also show correlations between PC scores and their IQ-adjusted versions, which were high for both PC1 ($\geq$ 0.89) and $PC1_{read}$ ($\geq$ 0.92) in all the datasets.

Familiality and heritability of PC scores

The assessment of sibling correlations for PC scores generally revealed moderate to high familialities and heritability estimates (Table 6), with UK-RD showing values lower than the other datasets. In this dataset PC1 showed a maximum heritability of 0.29 (vs values in the range [0.5-0.84] in the other datasets). Comparably, $PC1_{read}$ heritability in UK-RD was 0.27 (vs values in the range [0.47-0.75] in the other datasets). Nonetheless, these heritability estimates were significant at the $\alpha = 0.05$ level. This discrepancy was even more remarkable after IQ-adjustment, with UK-RD showing heritabilities of 0.16 and 0.15 for IQ-adjusted PC1 and IQ-adjusted $PC1_{read}$ (not significant at the $\alpha = 0.05$ level), whereas the other datasets reported heritabilities comparable to those of PC scores before IQ-adjustment (see Table 6).

4a)

Trait[a]	WRead	WSpell	PD (NWRead)	PD (PC)	PA	OC	NWR	PC1	PC1read
VIQ	0.592	0.462	0.433	0.376	0.371	0.373	0.448	0.534	0.558
PIQ	0.307	0.298	0.228	0.224	0.267	0.181	0.314	0.321	0.322

4b)

Trait[a]	WRead	WSpell	PD	PA	OC	PC1	PC1read
VIQ	0.463	0.32	0.311	0.357	0.372	0.429	0.419
PIQ	0.296	0.325	0.262	0.28	0.229	0.325	0.333

4c)

Trait[a]	WRead	WSpell	NWR	ELS	RLS	PC1	PC1read
VIQ	0.678	0.629	0.473	0.757	0.735	0.785	0.677
PIQ	0.376	0.372	0.234	0.385	0.453	0.463	0.388

4d)

Trait[a]	WRead	WSpell	PD (NWRead)	PD (PC)	PA	OC	NWR	PC1	PC1read
VIQ	0.596	0.434	0.489	0.334	0.433	0.323	0.135	0.582	0.572
PIQ	0.184	0.267	0.101	0.048	0.204	0.114	0.192	0.214	0.25

Table 4. Pairwise trait correlations of reading and language traits with IQ measures in **a)** CLDRC-RD, **b)** UK-RD, **c)** SLIC, and **d)** CLDRC-ADHD datasets. These were computed separately within each dataset, as the median Pearson's correlation coefficient over 100 repeat random samplings of one individual from each independent sibship.
[a] Legend: WRead = word reading; WSpell = word spelling; PD = phonological decoding (NWRead = nonword reading and PC = phonological choice); PA = phoneme awareness; OC = orthographic coding; NWR = nonword repetition; ELS/RLS = expressive/receptive language score; PC1 = first principal component derived from all the reading and language measures available in each dataset; $PC1_{read}$ = first principal component derived from word reading and spelling only; VIQ = verbal IQ; PIQ = performance IQ.

5a)

PC scores	PC1	PC1read	IQadjPC1	IQadjPC1read
PC1	1	0.925	0.941	0.865
PC1read		1	0.86	0.939
IQadjPC1			1	0.917
IQadjPC1read				1

5b)

PC scores	PC1	PC1read	IQadjPC1	IQadjPC1read
PC1	1	0.947	0.948	0.887
PC1read		1	0.889	0.942
IQadjPC1			1	0.939
IQadjPC1read				1

5c)

PC scores	PC1	PC1read	IQadjPC1	IQadjPC1read
PC1	1	0.914	0.893	0.806
PC1read		1	0.821	0.919
IQadjPC1			1	0.894
IQadjPC1read				1

5d)

PC scores	PC1	PC1read	IQadjPC1	IQadjPC1read
PC1	1	0.917	0.977	0.892
PC1read		1	0.882	0.968
IQadjPC1			1	0.913
IQadjPC1read				1

Table 5. Pairwise trait correlations of principal component scores in **a)** CLDRC-RD, **b)** UK-RD, **c)** SLIC, and **d)** CLDRC-ADHD datasets. These were computed separately within each dataset, as the median Pearson's correlation coefficient over 100 repeat random samplings of one individual from each independent sibship.
[a] Legend: PC1 = first principal component derived from all the reading and language measures available in each dataset; $PC1_{read}$ = first principal component derived from word reading and spelling only; IQadjPC1/IQadjPC1$_{read}$ = PC1/$PC1_{read}$ adjusted for performance IQ.

PC score[a]	CLDRC-RD	UK-RD	SLIC	CLDRC-ADHD
PC1	0.251 (0.5)	0.143 (0.29)	0.293 (0.59)	0.42 (0.84)
PC1read	0.233 (0.47)	0.137 (0.27)	0.261 (0.52)	0.373 (0.75)
IQadjPC1	0.267 (0.53)	0.081* (0.16)	0.325 (0.65)	0.438 (0.88)
IQadjPC1read	0.241 (0.48)	0.075* (0.15)	0.285 (0.57)	0.405 (0.81)

Table 6. Familiality of principal component scores in the datasets, and corresponding maximum heritability estimates in brackets. Familiality was computed as the median sibling correlation, over 100 random samplings, of one sibling pair from each family, separately in each dataset. Maximum heritability estimates were calculated as twice the corresponding familiality parameters, and represent both the contribution of additive genetic factors (i.e. heritability) and the potential contribution of shared environmental factors to the phenotypic variance of PC scores. All the familialities were significant at the $\alpha = 0.05$ level, except for the values labelled with "*".
[a] Legend: PC1 = first principal component derived from all the reading and language measures available in each dataset; $PC1_{read}$ = first principal component derived from word reading and spelling only; IQadjPC1/IQadj$PC1_{read}$ = PC1/$PC1_{read}$ adjusted for performance IQ.

Discussion

In the present chapter, we analyzed the reading and language traits available for our GWAS meta-analysis, statistically elaborated these measures, and derived component scores representing common variance in these traits within each dataset. This approach was suggested mainly by the moderate to high cross-phenotypic correlations detected among most of the reading and language traits available, supporting the hypothesis of a substantial shared variance in these traits and justifying the extraction of a First Principal Component (PC1) score. Additional reasons for analyzing a single trait representing all the reading and language traits available were the aim of reducing the number of traits to test for association -and therefore the correction for multiple testing of significance levels in the GWASMA (Chapter 3)- and the high computational load implied by multivariate association analyses in a GWA context, which made this kind of analysis unfeasible for a high number of SNPs.

In spite of the phenotypic heterogeneity of our datasets, PC1 can be considered broadly comparable across datasets for a number of reasons.

First, PC1 captured the majority of the common variance across the reading and language traits within each dataset (52-75%), with the second principal component (PC2) representing a proportion of common phenotypic variance from four- to six-fold lower. The loadings of the individual traits on PC1 scores were generally high and comparable in all the datasets.

Second, PC1 was strongly correlated with $PC1_{read}$, i.e. the First Principal Component derived only from word reading and spelling, which were the only two measures available in all of the datasets and provided the closest phenotype matching possible across datasets. As a

confirmation, the loadings of these two traits on $PC1_{read}$ were high and similar across the different datasets. These pieces of evidence further corroborated the comparability of our main trait of interest, namely PC1.

Third, PC1 was not affected by dropping/adding one or more traits in its computation, as revealed by the high correlations between the "original" PC1 scores, i.e. the final scores which were later used in the GWASMA, and the "experimental" PC1 scores, computed after adding or removing one or more reading/language traits from the PCA. This was assessed both in UK-RD and in SLIC, while no correlations were assessed in CLDRC datasets since all the measures available had low missing rates. High correlations between "original" and "experimental" PC1s suggested high reliability and robustness of these component scores, and allowed us to select the traits to be involved in the computation of the final PC1 scores without the concern of heavily affecting them.

The role of IQ in the etiology of RD/SLI and, more in general, in reading/language capacities is still debated: in addition to a substantial component of phenotypic variance shared between reading and language skills, but not with IQ, increasing evidence suggests the existence of phenotypic variance common to reading, language and general cognitive abilities (Bishop & Snowling, 2004; Pennington & Bishop, 2009). In line with this evidence, our reading and language traits generally showed moderate correlations with IQs, both at the individual trait and at the principal component level. This is consistent with the view that some genetic effects on reading and language may be pleiotropic for IQ (Bishop & Snowling, 2004; Pennington & Bishop, 2009). Furthermore, it underlined the need to analyse IQ-adjusted versions of our PC scores, namely IQ-adjusted PC1 and IQ-adjusted $PC1_{read}$. These would have later been useful to detect genetic variants with pleiotropic effects on reading and language skills but not on general (nonverbal) cognitive abilities (see Chapter 3, 4, 5).

Finally, we assessed heritability of our reading and language scores by calculating sibling correlation of our principal component scores, and then multiplying these values by two. This kind of familiality-based estimate may represent an overestimation of heritability, as it also includes the fraction of phenotypic variance attributable to shared environmental factors. However, it was still useful to compare our results with previous heritability estimates of reading and language traits and deficits (see *Introduction* for an overview).

Heritability of PC1 and $PC1_{read}$ was moderate to high in all the datasets, with UK-RD showing lower values than the other datasets, and this discrepancy was even more

pronounced for IQ-adjusted PC scores. This inconsistency may be explained through the differential recruitment of RD probands in UK-RD, where cases were collected through several reading clinics and hospitals across the UK and the recruitment was initially focused on sibling pairs where both members were very severe RD cases. The reduced variance in their reading/language scores may have affected the sibling correlations in this dataset, although other studies involving subsets of this dataset have reported higher familialities and heritabilities both for univariate traits (Francks et al., 2003) and for the first principal component score derived from them (Francks, 2001).

It has been hypothesized that a substantial variance associated with environmental variation may contribute to underestimate heritability in a dataset (Bishop & Snowling, 2004). In other words, if the quality of instruction, motivation to learn, socio economic status or other environmental factors relevant to reading and language learning vary widely within a sample, then heritability estimates may be lower than if all children were exposed to a more uniform environment. It is possible that this factor may have biased the heritability estimates in UK-RD.

By contrast, maximum heritabilities in both CLDRC subsets were concordant with previous heritability estimates made in the Colorado dataset through twin based studies, both for single univariate reading/language traits (Gayán & Olson 2001; 2003) and for composite (Friend et al., 2008) or principal component scores derived from them (Francks 2001). These results are also in line with the heritabilities reported by Francks et al. (2003) for a subset of the UK-RD dataset. Similarly, heritabilities obtained in SLIC were comparable with the SLI heritabilities reported by previous twin studies (Bishop et al 1995; 1999; Tomblin and Buckwalter, 1998; Viding et al., 2004), and with the heritability observed for typical phenotypic markers of SLI, namely nonword repetition (Bishop et al., 1999) and receptive language (Dale et al., 2010).

Overall, the findings of this chapter support the existence of a substantial shared variance between reading and language traits, of which PC1 can be considered an appropriate proxy measure. PC1 exhibited moderate to high heritabilities in the present work, which suggests an important genetic influence on the common variance in reading and language traits, in line with previous investigations on these phenotypes. These assumptions are consistent with the hypothesis that reading and language disorders are due at least in part to the same cognitive deficits (Bishop & Snowling, 2004; Pennington & Bishop, 2009), and with the view that the same biological/genetic bases may subserve different cognitive functions through pleiotropic effects, as postulated by the "generalist gene" hypothesis (Plomin & Kovas, 2005; Kovas &

Plomin, 2006). PC1 shows the advantage of being a broadly comparable trait across phenotypically heterogeneous datasets, allowing to overcome issues of heterogeneity of recruitment and assessment of different cohorts, a limitation often found in meta-analysis studies. All these elements support the suitability of PC scores to genetic association analyses and justify their use in the search for genetic variants with pleiotropic effects on reading and language traits through a GWAS meta-analysis, which will be described in Chapter 3. Finally, these findings provide an interesting perspective of the behavioral genetics of reading and language, focusing the attention on the common variance between reading and language abilities, rather than on individual skills and deficits affecting specific cognitive domains.

Supplementary Material

- *S1:* Description of the reading and language traits assessed in each dataset, including information on the statistical elaboration that they underwent.
- *S2:* Number of measures available for each reading/language trait in the different datasets. Additional notes on the composition of PC1 scores.

References

Almasy, L. & Blangero, J. (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet*, **62**, 1198-1211.

Bishop, D.V.M., Bishop, S.J., Bright, P., James, C., Delaney, T. & Tallal, P. (1999) Different Origin of Auditory and Phonological Processing Problems in Children With Language Impairment: Evidence From a Twin Study. *J Speech Lang Hear Res,* **42,** 155.

Bishop, D.V.M. & Hayiou-Thomas, M.E. (2008) Heritability of specific language impairment depends on diagnostic criteria. *Genes Brain Behav,* **7,** 365-372.

Bishop, D.V.M., North, T. & Donlan, C. (1995) Genetic basis of Specific Language Impairment: evidence from a twin study. *Dev Med Child Neurol,* **37,** 56-71.

Bishop, D.V.M. & Snowling, M.J. (2004) Developmental Dyslexia and Specific Language Impairment: Same or Different? *Psychol Bull,* **130,** 858-886.

Carrion-Castillo, A., Franke, B. & Fisher, S.E. (2013) Molecular Genetics of Dyslexia: An Overview. *Dyslexia,* **19,** 214-240.

Castles, A. & Coltheart, M. (1993) Varieties of developmental dyslexia. *Cognition,* **47,** 149-180.

Coltheart, M. & Leahy, J. (1996) Assessment of lexical and non-lexical reading abilities in children: some normative data. *Aust J Psychol* **48,** 136-140.

Compton, D.L., DeFries, J.C. & Olson, R.K. (2001) Are RAN- and phonological awareness-deficits additive in children with reading disabilities? *Dyslexia,* **7,** 125-149.

Dale, P.S., Harlaar, N., Hayiou-Thomas, M.E. & Plomin, R. (2010) The Etiology of Diverse Receptive Language Skills at 12 Years. *J Speech Lang Hear Res*, **53**, 982-992.

DeFries, J.C., Filipek, P.A., Fulker, D.W., Olson, R.K., Pennington, B.F., Smith, S.D. & Wise, B.W. (1997) Colorado Learning Disabilities Research Center. *Learn Disabil Multidisciplinary J,* **8,** 7-19.

DeFries, J.C. & Fulker, D.W. (1985) Multiple regression analysis of twin data. *Behav Genet,* **15,** 467-473.

DeFries, J.C. & Gillis, J.J. (1193) In Plomin, R. & McClearn, G. (eds), *Nature, Nurture, and Psychology.* American Psychiatric Association, Washington DC, pp. 121-145.

Dunn, L.M. & Markwardt, F.C. (1970) *Examiner's Manual: Peabody Individual Achievement Test,* American Guidance Service: Circle Pines, MN.

Eicher, J.D., Powers, N.R., Miller, L.L., Akshoomoff, N., Amaral, D.G., Bloss, C.S., Libiger, O., Schork, N.J., Darst, B.F., Casey, B.J., Chang, L., Ernst, T., Frazier, J., Kaufmann, W.E., Keating, B., Kenet, T., Kennedy, D., Mostofsky, S., Murray, S.S., Sowell, E.R., Bartsch, H., Kuperman, J.M., Brown, T.T., Hagler, D.J., Jr., Dale, A.M., Jernigan, T.L., Pourcain, B.S., Smith, G.D., Ring, S.M. & Gruen, J.R. (2013) Genome-Wide Association Study of Shared Components of Reading Disability and Language Impairment. *Genes Brain Behav*, **12**, 792-801.

Elliot, C.D., Murray, D.J. & Pearson, L.S. (1979) *The British ability scales,* NFER, Slough, UK.

Falcaro, M., Pickles, A., Newbury, D.F., Addis, L., Banfield, E., Fisher, S.E., Monaco, A.P., Simkin, Z. & Conti-Ramsden, G. (2008) Genetic and phenotypic effects of phonological short-term memory and grammatical morphology in specific language impairment. *Genes Brain Behav,* **7,** 393-402.

Field, A. (2005) *Discovering statistics using SPSS,* SAGE Publications Ltd, London.

Fisher, S.E. & DeFries, J.C. (2002) Developmental dyslexia: genetic dissection of a complex cognitive trait. *Nat Rev Neurosci,* **3,** 767-780.

Francks, C. (2001) Mapping of genes causing dyslexia susceptibility. University of Oxford.

Francks, C., Fisher, S.E., Marlow, A.J., MacPhie, I.L., Taylor, K.E., Richardson, A.J., Stein, J.F. & Monaco, A.P. (2003) Familial and genetic effects on motor coordination, laterality, and reading-related cognition. *Am J Psychiatry,* **160,** 1970-1977.

Francks, C., Paracchini, S., Smith, S.D., Richardson, A.J., Scerri, T.S., Cardon, L.R., Marlow, A.J., MacPhie, I.L., Walter, J., Pennington, B.F., Fisher, S.E., Olson, R.K., DeFries, J.C., Stein, J.F. & Monaco, A.P. (2004) A 77-Kilobase Region of Chromosome 6p22.2 Is Associated with Dyslexia in Families From the United Kingdom and From the United States. *The Am J Hum Genet,* **75,** 1046-1058.

Frederickson, N. (1995) *Phonological assessment battery,* London: Educational Psychology Publishing.

Friend, A. & Olson, R.K. (2010) Phonological Spelling and Reading Deficits in Children with Spelling

Disabilities. *Sci Stud Read,* **12,** 90-105.

Gallagher, A. & Frederickson, N. (1995) The phonological assessment battery (PhAB): an initial assessment of its theoretical and practical utility. *Educ Child Psychol* **12,** 53-67.

Gathercole, S., Willis, C., Baddeley, A. & Emslie, H. (1994) The children's test of nonword repetition: a test of phonological working memory. *Memory,* **2,** 103 - 107.

Gayan, J. & Olson, R.K. (2001) Genetic and Environmental Influences on Orthographic and Phonological Skills in Children With Reading Disabilities. *Dev Neuropsychol,* **20,** 483-507.

Gayán, J. & Olson, R.K. (2003) Genetic and environmental influences on individual differences in printed word recognition. *J Exp Child Psychol,* **84,** 97-123.

Harlaar, N., Hayiou-Thomas, M.E., Dale, P.S. & Plomin, R. (2008) Why Do Preschool Language Abilities Correlate With Later Reading? A Twin Study. *J Speech Lang Hear Res,* **51,** 688-705.

Jastak, S. & Wilkinson, G.S. (1984) *Wide Range Achievement Test - Revised (WRAT-R).* San Antonio, TX: The Psychological Corporation.

Kovas, Y. & Plomin, R. (2006) Generalist genes: implications for the cognitive sciences. *Trends Cogn Sci,* **10,** 198-203.

Logan, J., Petrill, S., Flax, J., Justice, L., Hou, L., Bassett, A., Tallal, P., Brzustowicz, L. & Bartlett, C. (2011) Genetic Covariation Underlying Reading, Language and Related Measures in a Sample Selected for Specific Language Impairment. *Behav Genet,* **41,** 651-659.

Luciano, M., Evans, D.M., Hansell, N.K., Medland, S.E., Montgomery, G.W., Martin, N.G., Wright, M.J. & Bates, T.C. (2013) A genome-wide association study for reading and language abilities in two population cohorts. *Genes Brain Behav,* **12,** 645-652.

McArthur, G.M., Hogben, J.H., Edwards, V.T., Heath, S.M. & Mengler, E.D. (2000) On the "specifics" of specific reading disability and specific language impairment. *J Child Psychol Psychiatry,* **41,** 869-874.

McGrath, L.M., Pennington, B.F., Shanahan, M.A., Santerre-Lemmon, L.E., Barnard, H.D., Willcutt, E.G., DeFries, J.C. & Olson, R.K. (2011) A multiple deficit model of reading disability and attention-deficit/hyperactivity disorder: searching for shared cognitive deficits. *J Child Psychol Psychiatry*, **52**, 547-557.

Mueller, K.L. (2012) Causation, correlation, or confound? What the comorbidity of language impairment and ADHD can tell us about the etiology of these disorders. University of Iowa.

Newbury, D., Winchester, L., Addis, L., Paracchini, S., Buckingham, L., Clark, A., Cohen, W., Cowie, H., Dworzynski, K., Everitt, A., Goodyer, I., Hennessy, E., Kindley, A., Miller, L., Nasir, J., O'Hare, A., Shaw, D., Simkin, Z., Simonoff, E., Slonims, V., Watson, J., Ragoussis, J., Fisher, S., Seckl, J., Helms, P., Bolton, P., Pickles, A., Conti-Ramsden, G., Baird, G., Bishop, D. & Monaco, A. (2009) CMIP and ATP2C2 modulate phonological short-term memory in language impairment. *Am J Hum Genet,* **85,** 264 - 272.

Newbury, D.F. & Monaco, A.P. (2010) Genetic advances in the study of speech and language disorders. *Neuron,* **68,** 309-320.

Olson, R., Wise, B., Conners, F., Rack, J. and Fulker, D. (1989) Specific deficits in component reading and language skills: genetic and environmental influences. *J Learn Disabil,* **22,** 339–348.

Olson, R., Forsberg, H. & Wise, B. (1994a) Genes, environment, and the development of orthographic skills. In Berninger, V.W. (ed), *The Varieties of Orthographic Knowledge I: Theoretical and Developmental Issues.* Kluwer Academics: Dordrecht, The Netherlands, pp. 27–71.

Olson, R., Forsberg, H., Wise, B. & Rack, J. (1994b) Measurement of word recognition, orthographic, and phonological skills. In Lyon, O.R. (ed), *Frames of Reference for the Assessment of Learning Disabilities: New Views on Measurement Issues.* Paul H. Brookes Publishing Co, Baltimore, MD, pp. 243-278.

Pennington, B.F. (2006) From single to multiple deficit models of developmental disorders. *Cognition,* **101,** 385–413.

Pennington, B.F. & Bishop, D.V. (2009) Relations among speech, language, and reading disorders. *Annu Rev Psych,* **60,** 283-306.

Plomin, R. & Kovas, Y. (2005) Generalist Genes and Learning Disabilities. *Psychol Bull,* **131,** 592-617.

Raskind, W.H., Peter, B., Richards, T.L., Eckert, M.A. & Berninger, V.W. (2013) The genetics of reading

disabilities: from phenotypes to candidate genes. *Front Psychol,* **3**, 601.

Rust, J., Golombok, S. & Trickey, G. (1993) *Wechsler objective reading dimensions,* Psychological Corporation, Sidcup, Kent, UK.

Semel, E.M., Wiig, E.H. & Secord, W. (1992) *Clinical Evaluation of Language Fundamentals–Revised,* San Antonio: Psychological Corporation.

Snowling, M., Bishop, D.V. & Stothard, S.E. (2000) Is preschool language impairment a risk factor for dyslexia in adolescence? *J Child Psychol Psychiatry,* **41,** 587–600.

The SLI Consortium (2002) A Genomewide Scan Identifies Two Novel Loci Involved in Specific Language Impairment. *Am J Hum Genet,* **70,** 384-398.

The SLI Consortium (2004) Highly significant linkage to the SLI1 locus in an expanded sample of individuals affected by specific language impairment. *Am J Hum Genet,* **74,** 1225-1238.

Tomblin, J.B. & Buckwalter, P.R. (1998) Heritability of Poor Language Achievement Among Twins. *J Speech Lang Hear Res,* **41,** 188-199.

Viding, E., Spinath, F.M., Price, T.S., Bishop, D.V.M., Dale, P.S. & Plomin, R. (2004) Genetic and environmental influence on language impairment in 4-year-old same-sex and opposite-sex twins. *J Child Psychol Psychiatry*, **45**, 315-325.

Wechsler, D. (1974) *Manual for the Wechsler Intelligence Scale for Children - Revised,* New York, NY: The Psychological Corporation.

Wechsler, D. (1981) *Manual for the Wechsler Adult Intelligence Scale - Revised,* Psychological Corporation, New York.

Wechsler, D., Golombok, S. & Rust, J. (1992) *WISC-IIIUK: Wechsler Intelligence Scale for Children: UK manual, third edition.,* Sidcup, UK: The Psycological Corporation.

Willcutt, E.G., Betjemann, R.S., McGrath, L.M., Chhabildas, N.A., Olson, R.K., DeFries, J.C. & Pennington, B.F. (2010) Etiology and neuropsychology of comorbidity between RD and ADHD: the case for multiple-deficit models. *Cortex*, **46**, 1345-1361.

Willcutt, E.G., Pennington, B.F., Olson, R.K., Chhabildas, N. & Hulslander, J. (2005) Neuropsychological analyses of comorbidity between reading disability and attention deficit hyperactivity disorder: in search of the common deficit. *Dev Neuropsychol,* **27**, 35-78.

S1: Description of reading and language traits

Trait	Test[a]	Test description[b]	Statistical elaboration[c]
WRead	British Ability Scale (BAS)/Wide Range Achievement Test-Revised (WRAT-R)[1,2]	Reading aloud a series of real words presented on a card	A, S, R
WSpell	BAS/WRAT-R[1,2]	Writing words that are dictated by the test administrator	A, S, R
PD	Castles & Coltheart (C&C)[3,4] Nonword reading	Reading aloud nonsense words of increasing difficulty, according to English grapheme-phoneme conversion rules	A, S, R
PA	Spoonerism test[5,6]	Simple phoneme deletion and substitution (e.g. replace the first sound in *dog* with \l\ to make *log*) Complex phoneme deletion and substitution Spoonerism (swapping the first sounds of two words, e.g. from *spoon, dog* to *doon, spog*)	A, S, R
OC	C&C[3,4] Irregular word reading	Reading aloud irregular words of increasing difficulty (i.e. words whose pronunciation does not follow the English grapheme-phoneme conversion rules, e.g. *yacht*)	A, S, R
vIQ	BAS/ Wechsler Adult Intelligence Scale – Revised (WAIS-R)[1,7]	Similarities subtest only (explaining how two/three words are similar or go together)	A, S, R
pIQ	BAS[1]	Matrices subtest only (predicting missing components of increasingly complex matrices containing abstract symbols)	A, S

Table S1a. Language/reading-related traits available in the UK-RD dataset. [a] Superscript numbers after each test indicate the initial reference for it (where further details on the test can be found): 1. Elliot et al., 1979; 2.Jastak & Wilkinson, 1984; 3.Castles & Coltheart 1993; 4.Coltheart & Leahy 1996; 5.Gallagher & Frederickson 1995; 6.Frederickson 1995; 7.Wechsler 1981.[b] Where more than one battery is administered, the total score is usually computed as a sum of the raw scores from each subtest. [c] Legend of trait adjustments: A= age-adjusted; S= standardized against the normative mean of the population of reference; R= further rank-normalized (using Blom's formula) because the trait distribution after standardization differed from normality (Shapiro-Wilk test p-val < 0.05).

Trait	Test[a]	Test description[b]	Statistical elaboration[c]
WRead	Wechsler Objectives of Reading Dimensions (WORD)[1]	Reading single real words of increasing difficulty	A, S, R
WSpell	WORD[1]	Spelling of single real words	A
NWR	Gathercole & Baddeley[2]	Repeating tape-recorded nonsense words of increasing length and complexity	A, S, R
ELS	Clinical Evaluation of Language Fundamentals Revised (CELF-R)[3]	Formulating sentences (formulating sentences about visual stimuli using a targeted word or phrase) Recalling sentences (imitating sentences presented by the examiner) Sentence assembly (producing two semantically/grammatically correct sentences from visually and orally presented words/groups of words)	A, S, R
RLS	CELF-R[3]	Oral directions (pointing to pictured objects in response to oral directions) Semantic relations (listening to a sentence and selecting the two choices that answer a target question, out of four possible answers) Word classes (choosing two related words and describing their relationship)	A, S, R
vIQ	Wechsler Intelligence Scale for Children (WISC)/WAIS[4]	Arithmetic (solving orally administered arithmetic word problems) Comprehension (explaining situations, actions, or activities that the examinee is expected to be familiar with) Digit span (reciting a sequence of digits presented by the examiner by recalling them in the same/reverse order) Information (general cultural knowledge test) Similarities (explaining how two words are alike/similar) Vocabulary (defining a provided word)	A
pIQ	WISC/WAIS[4]	Block design (arranging blocks to duplicate a given image/design) Coding (marking rows of shapes with different lines/transcribing symbols under digits, according to a given code) Object assembly (correctly assembling the parts that an object is divided into, like a puzzle) Picture arrangement (arranging a number of given pictures from left to right to tell the intended story) Picture completion (identifying the missing part in a series of pictures representing common objects)	A, S, R

Table S1b. Language/reading-related traits available in the SLIC dataset. [a] Superscript numbers after each test indicate the initial reference for it (where further details on the test can be found): 1.Rust et al., 1993; 2.Gathercole et al., 1994; 3.Semel et al., 1992; 4.Wechsler et al., 1992.[b] Where more than one battery is administered, the total score is usually computed as a sum of the raw scores from each subtest. [c] Legend of statistical elaborations: A= age-adjusted; S= standardized against the normative mean of the population of study, when required (Shapiro-Wilk test p-val < 0.05); R= further rank-normalized (using Blom's formula) because the trait distribution after standardization differed from normality (Shapiro-Wilk test p-val < 0.05).

Trait	Test[a]	Test description[b]	Statistical elaboration[c]
WRead	Peabody Individual Achievement Test (PIAT)[1]	Reading aloud in sequence single real words increasing in semantic and phonetic difficulty, until errors are made in 5 out of any 7 consecutive items (untimed)	C, A, S, R
	Timed oral reading[2,3]	Reading aloud a series of single real words within 2 seconds of their presentation, until errors are made in 10 out of any 20 consecutive items	
WSpell	PIAT[1]	Choosing the correct spelling of a series of real words (of increasing difficulty) orally presented, among four orthographically and often phonologically similar alternatives printed on a card (for each word), until errors are made in 5 out of 7 consecutive responses	A, S
PD	Oral Nonword Reading Task[2,3]	Reading aloud a series of single-syllable nonsense words (structure ranging from *vcv* to *cccvcv*) Reading aloud a series of two-syllables nonsense words	C, A, S, R
	Phonological Choice (Silent Nonword Reading Task)[2,3]	Choosing which of three nonsense words would sound like a real word if read aloud (for n triplets of nonwords)	A, S, R
PA	Phoneme Segmentation and Transposition Task[3]	Taking the first phoneme of a word, putting it at the end and add the sound */ay/* (for n words, e.g. *rope* → *ope-ray*)	C, A, S, R
	Phoneme Deletion Task[3]	Repeating nonwords within 2 seconds of their oral presentation, then removing a specified phoneme and pronouncing the resulting words within another 4 seconds (e.g. "say *prot*..now say *prot* without the */r/*" "*pot*")	
OC	Word-Pseudohomophone Choice[2,4]	Speeded forced-choice to distinguish a real word from a phonologically similar nonword (for n pairs of words-nonwords; e.g. rane vs. *rain*)	C, A, S, R
	Homophone Choice[2,4]	Selecting which of two homophones visually presented answers a question asked orally by the tester (for n pairs of words, e.g. "Which is a flower?" *rose* rows)	
NWR	Gathercole & Baddeley[5]	Repeating tape-recorded nonsense words of increasing length and complexity	A, S, R
vIQ	WISC-R/WAIS-R[6]	Comprehension (explaining situations, actions, or activities that the examinee is expected to be familiar with) Information (general cultural knowledge test) Similarities (explaining how two words are alike/similar) Vocabulary (defining a provided word)	None
pIQ	WISC-R/WAIS-R[6]	Block design (arranging blocks to duplicate a given image/design) Object assembly (correctly assembling the parts that an object is divided into, like a puzzle) Picture arrangement (arranging a number of given pictures from left to right to tell the intended story) Picture completion (identifying the missing part in a series of pictures representing common objects)	None

Table S1c. Language/reading-related traits available in the CLDRC dataset. [a] Superscript numbers after each test indicate the initial reference for it (where further details on the test can be found): 1. Dunn & Markwardt, 1970; 2. Olson et al., 1989; 3. Olson et al., 1994a; 4. Olson et al., 1994b; 5. Gathercole et al., 1994; 6. Wechsler, 1974. [b] Where more than one battery is administered, the total score is computed as a sum of the raw scores from each subtest (IQ measures), as an average of z-scores derived from accuracy scores (% of correct responses) and median correct reaction times of the two subtests (nonword reading), or as the arithmetic average of the raw scores from each subtest (all the other measures).[c] Legend of statistical elaborations: C= composite score; A= age-adjusted (score regressed against age and age^2); S= standardized against the normative mean of a control population; R= further rank-normalized (using Blom's formula) because the trait distribution after standardization differed from normality (Shapiro-Wilk test p-val < 0.05).

S2: Further details on measures available and PC1 computation

Trait[a]	CLDRC-RD (564)	UK-RD (958)	SLIC (498)	CLDRC-ADHD (163)
WRead	564	953	273	163
WSpell	558	925	271	159
PD	560; 555[b]	950		163; 163[b]
PA	557	601		163
OC	557	946		163
NWR	560		472	163
ELS			426	
RLS			429	
VIQ	564	942	359	163
PIQ	564	911	461	163
PC1	544	914	245	159
IQadjPC1	544	878	245	159

Table S2. Number of measures available for reading- and language-related traits in each dataset and final number of PC1 and IQ-adjusted PC1 measures available in the GWAS meta-analysis. Sample sizes of the datasets (after phenotype QC, described in the present chapter, and genotype QC, described in Chapter 3) are reported in the header row.

[a] Legend: WRead = word reading; WSpell = word spelling; PD = phonological decoding; PA = phoneme awareness; OC = orthographic coding; NWR = nonword repetition; ELS/RLS = expressive/receptive language score; VIQ/PIQ = verbal/performance IQ; PC1 = first principal component derived from all the reading and language measures available in each dataset; IQadjPC1 = PC1 adjusted for performance IQ. [b] Number of measures available for nonword reading and phonological choice, respectively.

Additional notes on the composition of PC1 in the different datasets.

In UK-RD, the "original" PC1 (i.e. the final score which would have been later used in the GWAS meta-analysis) was derived from word reading, spelling, phonological decoding and orthographic coding, but not from phoneme awareness. This trait was excluded due to the high number of missing measures (~ 350), which implied a decrease in the sample size of the GWASMA by more than 300 subjects. By contrast, the original PC1 in SLIC was extracted from all the reading and language measures available, namely word reading, spelling, expressive and receptive language and nonword repetition, although both language scores and word reading and spelling showed a relatively high number of missing measures (see Table S2). However, in this case the resulting reduction in the final number of PC1 measures available in the dataset was lower than in UK-RD: sample size decreased by less than 30 subjects when including language scores in the PCA, and by less than 170 subjects when including word reading and spelling. As we aimed at analyzing also $PC1_{read}$ scores and compare this analysis with PC1 analysis, in SLIC we decided to include all the reading and language measures available in PC1 computation (including word reading and spelling), in spite of a moderate reduction in the total sample size of the meta-analysis (< 10%).

In the CLDRC datasets, there were low missing rates for all the measures (see Table S2), therefore we decided to use all the reading and language traits available in PC1 computation. These included word reading, spelling, phonological decoding (both nonword reading and phonological choice), phoneme awareness, orthographic coding and nonword repetition.

Chapter 3: Genome-wide screening for DNA variants associated with reading and language traits

This chapter is based on

Gialluisi, A., Newbury, D.F., Wilcutt, E.G., Olson, R.K., DeFries, J.C., Brandler, W.M., Pennington, B.F., Smith, S.D., Scerri, T.S., Simpson, N.H., The SLI Consortium, Luciano, M., Evans, D.M., Bates, T.C., Stein, J.F., Talcott, J.B., Monaco, A.P., Paracchini, S., Francks, C. & Fisher, S.E. (2014) Genome-wide screening for DNA variants associated with reading and language traits. *Genes, Brain and Behavior*, 13(7), 686-701

Abstract

Reading and language abilities are heritable traits that are likely to share some genetic influences with each other. To identify pleiotropic genetic variants affecting these traits, we first performed a Genome-wide Association Scan (GWAS) meta-analysis using three richly characterised datasets comprising individuals with histories of reading or language problems, and their siblings. GWAS was performed in a total of 1,862 participants using the first principal component computed from several quantitative measures of reading- and language-related abilities, both before and after adjustment for performance IQ. We identified novel suggestive associations at the SNPs rs59197085 and rs5995177 (uncorrected p ~10^{-7} for each SNP), located respectively at the *CCDC136/FLNC* and *RBFOX2* genes. Each of these SNPs then showed evidence for effects across multiple reading and language traits in univariate association testing against the individual traits. *FLNC* encodes a structural protein involved in cytoskeleton remodelling, while *RBFOX2* is an important regulator of alternative splicing in neurons. The *CCDC136/FLNC* locus showed association with a comparable reading/language measure in an independent sample of 6,434 participants from the general population, although involving distinct alleles of the associated SNP. Our datasets will form an important part of on-going international efforts to identify genes contributing to reading and language skills.

Introduction

Reading disability (RD, also known as developmental dyslexia) refers to a significant difficulty in reading that cannot be explained by obvious causes, such as sensory impairments or lack of educational opportunity (Shaywitz et al., 1990). Specific Language Impairment (SLI) is diagnosed as an unexpected difficulty or delay in acquiring spoken language abilities, despite normal hearing and intelligence, and in absence of overt neurological deficits (Bishop, 1994). RD and SLI are among the most prevalent neurocognitive disorders of school-aged children, with prevalence ≈5-8% in many populations (Shaywitz et al., 1990; Tomblin et al., 1997). Both are complex disorders with moderate to high heritabilities (30-70%) as assessed by studies of families and twins (Barry et al. 2007; Fisher & DeFries, 2002).

RD and SLI display high comorbidity: 43% of SLI children are later diagnosed with RD and up to 55% of dyslexic children meet criteria for SLI (Snowling et al. 2000; McArthur et al. 2000). Moreover, RD and SLI show comorbidity with other neurodevelopmental traits including Attention Deficit Hyperactivity Disorder (ADHD) (Wilcutt et al. 2010; Pennington, 2006) and Speech Sound Disorders (SSD) (Newbury & Monaco, 2010; Pennington & Bishop, 2009). It is likely that these disorders arise due to some shared genetic/neurobiological mechanisms, as well as non-shared causal factors (Paracchini 2011; Newbury et al., 2011). A study of twins by Harlaar et al. (2008) indicated that an association between early language and later reading is underpinned by common environmental and genetic influences, and a family study by Logan et al. (2011) also found significant genetic correlations of reading and language measures.

Variants of several genes have previously been associated with RD, most notably *DYX1C1* (15q21, Taipale et al., 2003), *KIAA0319* and *DCDC2* (6p22, Francks et al., 2004; Cope et al., 2005; Meng et al., 2005), *MRPL19/GCFC2* (2p12, Anthoni et al., 2007) and *ROBO1* (3p12, Hannula-Jouppi et al., 2005; Bates et al.; 2011). Similarly, some loci have been implicated in SLI; variants in genes such as *CNTNAP2* (7q35, Vernes et al., 2008) and *CMIP* and *ATP2C2* (16q23-24, Newbury et al., 2009) show associations with quantitative traits in children with typical SLI, while rare mutations of *FOXP2* (7q31, Fisher and Scharff, 2009) cause a monogenic speech and language disorder. These genes were mostly identified through linkage analysis followed by either positional cloning or else targeted association mapping. Functional analyses suggest that some of these genes mediate important processes in central nervous system (CNS) development, such as neuronal migration, axonal guidance and neurite

outgrowth (Carrion-Castillo et al. 2013; Vernes et al., 2011; Poelmans et al., 2011). A subset of the candidate genes may contribute to both RD and SLI, again indicating a partial genetic overlap for these traits (Newbury et al., 2011; Scerri et al., 2011b; Bates et al., 2011). Crucially, an overwhelming majority of the heritable variance in reading and language skills is unexplained, and the molecular mechanisms that contribute to RD and SLI remain largely unknown (Peterson & Pennington, 2012; Newbury & Monaco, 2010).

Some of the genetic variation contributing to RD and SLI is likely to also impact on reading/language skills in the general population (Luciano et al., 2007; Paracchini et al, 2008; 2011; Whitehouse et al., 2011; Bates et al. 2011; Scerri et al., 2011b). To detect previously undiscovered associations of common genetic variants with reading and language skills, it is therefore appropriate to sample broad ranges of the trait distributions in study datasets, while screening over the entire genome.

In recent years a small number of studies have tried to identify genes involved in reading and/or language through genome-wide association scanning (GWAS). An early GWAS for reading ability used DNA pooling of low versus high reading ability groups in ~1,500 7-year-old children, and a relatively low density SNP microarray with ~107,000 SNPs (Meaburn et al., 2008). The SNPs showing the largest allele frequency differences between low and high ability groups were further genotyped and tested in an additional sample of 4,258 children, with 10 SNPs finally showing nominally significant association with continuous variation in reading ability (Meaburn et al., 2008). A GWAS on mismatch negativity, which is a potential endophenotype of dyslexia derived from electroencephalography, has also been reported based on 386 dyslexic children, and showed replicable association of the SNP rs4234898 on 4q32 along with the haplotype rs4234898-rs11100040 (Roeske et al., 2011). These were shown to affect mRNA expression levels of *SLC2A3* (12p13), which codes for a neuronal glucose transporter, suggesting a possible role of glucose levels in memory performance necessary for speech perception in dyslexia (Roeske et al., 2011). More recently, a genome-wide linkage and association scan using ~133,000 SNPs, in 718 subjects from 101 dyslexia-affected families, reported a borderline significant association with dyslexia status at rs9313548, near *FGF18* (5q35.1), which is a gene involved in laminar positioning of cortical neurons during development (Field et al., 2013).

Two GWAS studies have directly attempted to identify shared genetic contributions to reading and language. Luciano et al. (2013), in a GWAS on quantitative reading and language traits in two population datasets (N~6,500), found the strongest association between

rs2192161, in the *ABCC13* pseudogene (21q11.2), and a nonword repetition measure ($p \sim 7x10^{-8}$), while rs4807927 (*DAZAP1*, 19p13.3) showed association with both word reading and a composite reading-spelling factor score ($p \sim 10^{-6}$ for both traits). In the same study, *CDC2L1, CDC2L2, LOC728661* (1p36.33) and *RCAN3* (1p36.11) showed significant gene-based associations with the reading-spelling factor (Luciano et al., 2013). A case-control GWAS using a relatively small number of RD (N=353), Language Impairment (N=163), and comorbid cases (N=174), in comparison to general population controls (N=4,117), identified nominally significant associations for the comorbid cases at rs12636438 and rs1679255 in *ZNF385D* (3p24.3) (Eicher et al., 2013). These SNPs also showed associations with a vocabulary measure and white matter volumes of brain fiber tracts previously implicated in language, in an independent dataset (Eicher et al., 2013).

In the present study we carried out a GWAS meta-analysis for genetic variants influencing reading and language abilities. We included three long-established datasets comprising children with reading or language problems, along with their siblings. This approach complemented other recent GWAS studies of reading/language performance (Luciano et al. 2013; Eicher et al., 2013) since it included continuous trait variance across a broad range of reading and language abilities, but also involved a pronounced enrichment for poor performance while not applying an arbitrary dichotomy between RD/SLI cases and controls.

Within each dataset we tested single nucleotide polymorphisms (SNPs), along with single base insertions/deletions (indels), for association with the first Principal Component (PC) derived from a range of reading- and language-related quantitative traits (see Chapter 2). We then meta-analyzed the GWAS results from the separate datasets, followed by gene- and pathway-level analysis, and we checked the most significant associations arising from our analysis within the GWAS results generated by Luciano et al. (2013).

Although we used PC-based analysis as a form of data reduction for the purposes of GWAS, we also investigated the two most significant SNP associations arising from our meta-analysis by using multivariate association modelling in each dataset, and by testing of these SNPs against the individual measures separately. This approach would help to understand the cross-phenotypic effects involved. In other words, the PC-based GWAS was used to identify potential genetic effects on shared variance between multiple reading and language measures, and then pleiotropy was investigated in more detail through univariate analysis and multivariate modelling, for individual SNPs implicated by the PC-based GWAS meta-analysis. In addition, in order to more closely match the trait measurement across all datasets

we repeated the GWAS and meta-analysis using the first PC of only single word reading and spelling ability, since these were the only two measures available in all datasets.

Some genetic effects on reading and language may be pleiotropic for IQ, whereas other effects may be largely or wholly independent of IQ (Bishop & Snowling, 2004; Pennington & Bishop, 2009). To detect the latter type of effect it is advantageous to remove the shared variance with IQ that is present in measures of reading and language, prior to association testing. We therefore performed our GWAS analyses both with and without IQ-adjustment of the reading and language measures. In addition, Luciano et al. (2013) analysed only IQ-adjusted data, so that for cross-comparing of results an IQ-adjustment was desirable to include in the present study.

Subjects and Methods

Datasets

Below the datasets involved in the study are briefly described. Further details are reported in Chapter 2 (see *Subjects and Methods* section).

UK-RD

This dataset comprised children diagnosed with RD, and their siblings, collected in several specialized clinics in the United Kingdom. The total number of participants was 983, mean age 11.7 years, age range 5-31, from 608 independent nuclear families. All children, regardless of diagnosis, were administered psychometric tests of reading- and language-related abilities, as well as assessments of verbal and non-verbal IQ (details in Chapter 2). A subset of this dataset has been analyzed in previous studies on reading (Becker et al., 2013) and handedness traits (Scerri et al., 2011a; Brandler et al., 2013), but no GWAS of reading/language-related traits has previously been reported.

SLIC

The SLI Consortium dataset comprised children affected by SLI, along with their siblings, recruited from five specialized centres across the UK. All children in this sample were assessed for a number of reading- and language-related traits regardless of their language

ability. For this study we obtained genome-wide genotype data for 548 participants, mean age 10 years, age range 5-19, from 288 independent nuclear families. The SLIC dataset has been used for prior linkage studies (SLIC 2002; 2004; Falcaro et al., 2008), and targeted candidate gene analyses (Vernes et al., 2008; Newbury et al., 2009). More recently, it has been used for investigating copy number variants (Ceroni et al., 2014), identification of chromosomal abnormalities (Simpson et al., 2014) and in a genome-wide search for parent-of-origin effects on SLI (Nudel et al. 2014). However, no GWAS for continuous language and reading scores has yet been reported for this (or any other) SLI sample.

CLDRC

The Colorado Learning Disabilities Research Centre (CLDRC) dataset was derived from an ongoing study on the etiology of learning disabilities run in 27 school districts in Colorado, USA (DeFries et al., 1997; Willcutt et al., 2005). Twins were initially recruited based on a school report of RD, ADHD or other learning disabilities along with their additional co-siblings; they were then administered a number of psychometric tests for several learning-related skills, and DNA was collected for genetic studies. The sample of twins and siblings available for this study comprised 749 participants in total, mean age 11.7 years, age range 8-19, from 343 unrelated twinships/sibships. Of these, 266 of the twinships/sibships (585 participants) were originally recruited via a proband with a history of RD, and 77 of the twinships/sibships (164 participants) were originally recruited via a proband with a history of ADHD. We analyzed these two subsets separately for GWAS before meta-analyzing the results together with those from the other datasets listed above. As in Chapter 2, the two subsets are indicated as CLDRC-RD and CLDRC-ADHD. As for the other datasets, no prior GWAS has been reported.

Genotype data generation, quality control (QC) and imputation

DNA was extracted from whole blood or buccal swab samples and prepared for genotyping using standard protocols. Genome-wide genotype data were generated for each dataset using Illumina® SNP arrays. These were the HumanHap 550k for a first genotyping wave of 200 subjects from UK-RD, and the Human OmniExpress (730k SNPs) for SLIC, CLDRC and the remaining UK-RD samples. Data were processed using Illumina's BeadStudio®/GenomeStudio® software, following the manufacturer's guidelines. All

datasets then underwent a first round of quality control, using functions in the software PLINK v1.07 (Purcell et al., 2007; http://pngu.mgh.harvard.edu/~purcell/plink/), in which all SNPs deviating from Hardy-Weinberg Equilibrium (HWE, $p < 1x10^{-6}$), with Minor Allele Frequency (MAF) < 1%, and call frequency < 99%, were filtered out. In addition, samples were excluded if they showed inconsistencies in genome-wide identity-by-descent sharing with their siblings and unrelated individuals, or sex mismatches, or call rates < 98%. Multi-Dimensional Scaling (MDS) analysis of genome-wide genotype data was used to identify any subjects that did not cluster together with the majority of the dataset, and these were discarded, as were any outliers for genome-wide homozygosity. These QC steps were followed by genotype phasing using MACH v1.0 (Li et al., 2010; http://www.sph.umich.edu/csg/abecasis/MACH/index.html) and imputation of SNPs and single-base indels using Minimac (Howie et al., 2012; http://genome.sph.umich.edu/wiki/Minimac), with the 1000 Genomes Project reference dataset (GIANT all populations panel, Phase 1, v3; The 1000 Genomes Project Consortium, 2012; http://www.1000genomes.org). We excluded poorly imputed polymorphisms (with $r^2 < 0.3$), and deleted individual genotypes with imputation quality scores < 0.9. A final quality control procedure was then run on the imputed data, using PLINK, in which we discarded SNPs with HWE $p < 5x10^{-6}$, MAF < 1%, and call frequency < 95%. Key features of the QC are shown in Table 1. Further details are reported in *Supplementary Material S1*.

At the end of the genotype QC process, we had data for 959 participants and 6,190,549 polymorphisms in UK-RD, 729 participants and 6,427,000 polymorphisms in CLDRC, and 502 participants and 6,240,842 polymorphisms in SLIC, with 5,518,496 polymorphisms shared across all three datasets.

QC step	CLDRC (749)[b]	UK-RD (200+818)[c]	SLIC (548)
HWE $p < 1 \times 10^{-6}$ (SNPs)	57	12,631[d]; 191	54
MAF < 1% (SNPs)	74,770	23,467; 77,342	1,718
Call Freq < 99% (SNPs)	0 [e]	82,052; 0 [e]	72,043
Call Rate < 98% (samples)	0 [e]	3; 0 [e]	9
IBD sharing (samples)	11	1; 7	17
Sex mismatch (samples)	3	0; 8 [f]	13 [g]
Homozygosity outlier (samples)	6	1; 3	2
MDS outlier (samples)	0	0; 2	5
HWE $p < 5 \times 10^{-6}$ (SNPs) [a]	2,166	2,779	2,096
MAF < 1% (SNPs)[a]	3,640,742	1,980,500	3,260,639
Call Freq <95% (SNPs)[a]	1,729,493	1,704,412	1,766,376
Call Rate < 95%, MDS outliers, IBD sharing (samples) [a]	0	0	0
Passing QC	729 (6,427,200)	959 (6,190,549)	502 (6,240,842)

Table 1. Genotype quality control (QC) filters used, and number of samples/markers discarded at each step (see *Subjects and Methods* and *Supplementary Material S1* for details). Final number of samples (and SNPs in brackets) passing the genotype QC are reported in the bottom row. Note that these numbers do not also account for QC of the trait scores.

[a] After imputation QC. Before this step, imputed SNPs with $r^2 < 0.3$ were filtered out, and all the genotypes with quality score < 0.9 were set to missing. [b] Since *CLDRC-RD* and *CLDRC-ADHD* were processed together and drawn from the same population, we treated them as a single dataset in the genotype QC. [c] Since *UK-RD* samples had been genotyped on two different Illumina® platforms (see *Subjects and Methods*), the subsets were analyzed separately before imputation, and pre-imputation QC details are therefore reported for both the subsets (first genotyping wave with HumanHap 550k and second genotyping wave with Human OmniExpress). Note that 35 samples were genotyped on both of the arrays, and one of these samples showed inconsistent genotyping and was therefore discarded in both subsets. [d] The high number of SNPs discarded at this stage was due to the fact that no quality filter had been applied on this subset during genotype call process (see *Supplementary Material S1*). [e] In this case, SNPs with call frequency < 99% and samples with call rate < 98% had already been discarded during genotype call process (see *Supplementary Material S1*). [f] Includes 3 sex chromosome abnormalities carriers. [g] Includes 9 samples with sex chromosome abnormalities and 1 with X chromosome call rate < 95%.

Reading and language traits: principal component scores

The reading- and language-related traits that were assessed in the different datasets are reported in Chapter 2 (see Table 1 and Tables S1a, b, c), along with details on phenotypic quality control and First Principal Component scores computation. Briefly, reading and language traits had been previously age-adjusted according to normative data, and underwent a further rank-normalization when required, to attain normality of distributions within datasets. Phenotypic outliers for three or more trait scores were discarded (one participant in UK-RD and one in CLDRC-RD), as well as subjects with full scale IQ < 70 (one participant from CLDRC-RD, and four participants from SLIC). This left 564 subjects in CLDRC-RD, 958 in UK-RD, 498 in SLIC and 163 in CLDRC-ADHD. For these subjects, the First Principal Component from all of the language- and reading-related traits available (PC1) was

extracted through a Principal Component Analysis (PCA) within each dataset. PC1s represented a substantial proportion of the common variance among the reading and language traits in all the datasets, and presented a broad pattern of loadings across all the traits (see Chapter 2). A version of PC1 adjusted for performance IQ was also computed (IQ-adjusted PC1). The final sample size for PC1 and IQ-adjusted PC1 meta-analysis (all datasets combined) was 1,862 and 1,826, respectively (see Table 2 for details). Similarly, we derived a first principal component score within each dataset from only word reading and spelling ($PC1_{read}$, N=1,913), to provide the closest phenotype matching possible across datasets, and computed an IQ-adjusted version (IQ-adjusted $PC1_{read,}$ N=1,875). This trait presented even higher loadings of word reading and spelling in each dataset, and explained a high fraction of their common variance. Moreover, the correlation between PC1 and $PC1_{read}$ was high in each dataset, so that PC1 itself could also be regarded as highly comparable across datasets (see Chapter 2 for details). In the present study, we primarily focused on PC1 for our subsequent genetic analysis (below), because this would maximise the chance of identifying SNPs that affect variance shared between both reading and language measures. However, we also repeated GWAS meta-analysis using $PC1_{read}$ to provide a comparable analysis that would be minimally affected by the heterogeneity of available measures across datasets.

PC score	Description	CLDRC -RD (564)	UK-RD (958)	SLIC (498)	CLDRC -ADHD (163)
PC1	Common variance in reading and language skills	544	914	245	159
IQ-adjusted PC1	Common variance in reading and language skills, not shared with general (nonverbal) cognitive abilities	544	878	245	159
PC1read	Common variance in strictly reading-related skills	558	925	271	159
IQ-adjusted PC1read	Common variance in strictly reading-related skills, not shared with general (nonverbal) cognitive abilities	558	888	270	159

Table 2. Principal component (PC) scores meta-analyzed in the present study and sample sizes of single GWAS within each dataset. Sample sizes of the datasets after genotype and phenotype QC, but before PC extraction, are reported in the header row. Sample sizes involved in the PC1 and $PC1_{read}$ meta-analyses are generally lower as we excluded participants with at least one missing measure among the traits involved in principal component analysis (see Chapter 2 for details on PC scores extraction). In the present study, we primarily focused on the GWAS meta-analysis of PC1 and IQ-adjusted PC1, to detect genetic effects on the variance shared among reading and language measures. GWAS meta-analysis of $PC1_{read}$ scores was run to provide a supporting analysis that would be minimally affected by the heterogeneity of traits available across datasets.

Genetic association analyses

Sibling-pair GWAS

Sibling-based genome-wide association analyses were conducted using PC1 and $PC1_{read}$ scores separately within each dataset, both before and after IQ-adjustment, and using the 'total' association option of the QFAM function implemented in PLINK v1.07 (http://pngu.mgh.harvard.edu/~purcell/plink/; Purcell et al., 2007). This method tests for association at each SNP by regressing trait scores on genotypes in an additive linear model. To correct for non-independence of siblings, permutations were run (i.e. label-swapping of phenotypes/genotypes) to obtain empirical significance levels (further details in *Supplementary Material S1*).

GWAS Meta-Analysis (GWASMA)

The results from GWAS in the separate datasets were then meta-analysed together. This was implemented in the program METAL (http://www.sph.umich.edu/csg/abecasis/Metal/index.html; Willer et al. 2010). We chose an approach that does not assume equivalence of allelic effect sizes between datasets, which was appropriate given the heterogeneity of study recruitment and assessment. Put briefly, the GWAS meta-analysis tested each SNP for a genetic effect, across the contributing datasets, computing an overall z-score for that SNP determined by the p-value, the direction of the allelic effect on the quantitative trait, and the sample size of each study involved in the meta-analysis.

Gene-based analysis

The results of the GWASMA on PC1 were used as input for gene-based association analyses using VEGAS v0.8.27 (http://gump.qimr.edu.au/VEGAS/; Liu et al. 2010). This software performs association tests for ~18,000 autosomal genes, by assigning multiple SNPs to each individual gene according to their genomic locations, and then combining the evidence for association across all SNPs assigned to a given gene, while taking into account the linkage disequilibrium (LD) structure between SNPs. Each tested gene also included potentially regulatory regions located up to 50 kb beyond the 5'- and 3'-untranslated regions (UTRs). A

Bonferroni-corrected significance threshold was set at $p < 2.8\times10^{-6}$ to account for the number of genes tested (see *Supplementary Material S1* for details).

Pathway-based analysis

Finally, a pathway/network-based association analysis was run using the PC1 GWASMA results, with the program INRICH v1.0 (http://atgu.mgh.harvard.edu/inrich/started.html; Lee et al. 2012). This tool tests for an enrichment of association within predefined gene sets, through a permutation-based approach. We defined associated genomic intervals as those containing an individual association $p < 0.001$ in the GWASMA results. Gene boundaries were again defined as extending 50 kb beyond the 5'- and 3'-UTRs. Three candidate gene lists, based on the gene sets of the Gene Ontology Database (http://www.geneontology.org/), were tested for an enrichment of association. These represented three distinct neurobiological hypotheses on the etiology of reading and language disabilities (see *Discussion* section for further explanations): axon guidance (including all the GO sets containing the term "axon guidance"), neuronal migration (including all the GO sets containing the term "neuron migration") and steroid sex hormone biology (including all the GO sets containing the terms "steroid", "androgen", "estrogen", "progesterone" and "testosterone"). Further details on the analysis can be found in *Supplementary Material S1*.

Further analysis of top association signals

Effect sizes on different traits

We repeated the regressions of PC1 and IQ-adjusted PC1 on the genotypes of our two most significantly associated SNPs from GWAS meta-analysis, in an additive linear model, in order to conveniently obtain the regression r^2 as indicative measures of effect sizes. To generate measures unbiased by sample relatedness, regression r^2 were calculated in R (R core Team, 2013, http://www.r-project.org/) as the median r^2 over 100 repeat random samplings of one individual from each independent sibship, separately in each dataset.

We further investigated each of our top two association signals by running QFAM univariate association tests in PLINK v1.07 (Purcell et al., 2007) for each individual trait that was used in constructing PC1, and separately in each dataset. This analysis provided an initial assessment of pleiotropy for these loci. We also performed multivariate association analysis

for these two loci, in PLINK Multivariate v1.06 (https://genepi.qimr.edu.au/staff/manuelF/multivariate/main.html; Ferreira & Purcell, 2009), again separately in each dataset and using each of the reading/language traits that were used in constructing PC1. PLINK Multivariate extracts the linear combination of traits that explains the largest possible amount of covariance between the SNP and all of the traits. The loading produced for each trait represents its contribution to the multivariate association. MQFAM 'total' association was run, with adaptive permutations to adjust for sample relatedness (see *Supplementary Material S1* for details).

Assessment of top association signals in two additional datasets

Our two most significant association signals from PC1 meta-analysis were checked against published and unpublished results from the recent GWASMA of reading and language abilities reported by Luciano et al. (2013). This prior study analysed two population datasets, the Brisbane Adolescent Twin Sample (*BATS*) and the Avon Longitudinal Study of Parents and their Children (*ALSPAC*). *BATS* is a cohort of twins and their non-twin siblings recruited from ongoing studies of melanoma risk factors and cognition in an Australian population-based sample (Wright et al., 2001). Subjects had been administered psychometric tests assessing regular-word, irregular-word, and non-word reading, and spelling, together with the Schonell graded word reading test, and nonword repetition (see Luciano et al., 2013). *ALSPAC* is a longitudinal, population-based sample recruited from the county of Avon, UK (Boyd et al., 2013). The study website contains details of all the data available through a fully searchable data dictionary (http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary). Ethical approval was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Participants (all free of neurological/psychiatric conditions) had been tested for word reading, nonword reading, spelling and nonword repetition (see Luciano et al., 2013). BATS and ALSPAC had been genotyped using Illumina® 610k Quad Bead and HumanHap 550k Quad chips respectively and imputed using the HapMap Phase II CEU reference panel (NCBI build 36) (The International HapMap 3 Consortium, 2010). A total of 6,434 subjects (962 from BATS and 5,472 from ALSPAC) were meta-analysed by Luciano et al. (2013), for three different traits: word reading, nonword repetition and a composite/component score of reading and spelling (called hereafter the *reading-spelling factor*).

Results

GWAS meta-analysis

Table 3 describes the most significant associations from the meta-analyses on PC1 (N=1,862) and IQ-adjusted PC1 (N=1,826). Figure 1 shows genome-wide Manhattan Plots. QQ-plots revealed no evidence of population stratification affecting the meta-analysis results, nor of genome-wide significant associations (Figure S2a, b). The most significant association was observed for rs59197085 in PC1 and IQ-adjusted PC1 meta-analyses ($p = 3.86x10^{-7}$ for PC1, and $p = 3.01x10^{-7}$ for IQ-adjusted PC1; A/G, MAF ~ 8%). This SNP is located at 7q32.1, within *CCDC136* (coiled-coil domain containing 136, or *NAG6*) and ~10 kb upstream of *FLNC* (filamin C; Fig. S2c). The second most significantly associated region, before IQ-adjustment, was located on 22q12.3, SNP rs5995177 ($p = 5.01x10^{-7}$, A/G, MAF ~ 8%), within *RBFOX2* (RNA-binding protein, fox-1 homolog 2, also known as RNA-binding motif protein 9, or *RBM9*; Fig. S2d). The association was less significant after IQ-adjustment of PC1 ($p = 1.5x10^{-5}$), and this difference was not merely due to the loss of 36 subjects in the IQ-adjusted analysis (investigated by performing a repeat PC1 analysis in the same reduced set of subjects as were available for IQ-adjusted PC1, data not shown).Table S2a, b shows all SNPs with association $p < 1x10^{-5}$ in GWAS meta-analysis of PC1 or IQ-adjusted PC1. No genome-wide significant associations were observed in the GWAS in the individual datasets (data not shown).

The results of our complementary $PC1_{read}$ meta-analysis (*Supplementary Material S3*) were consistent with the PC1 meta-analysis, with rs59197085 and rs5995177 among the top suggestive associations ($p \sim 10^{-6}$). This was expected given the high correlations between PC1 and $PC1_{read}$ in each dataset (all correlations ≥ 0.9, see Chapter 2).

Effect sizes and profiles of top associations

rs59197085 (*CCDC136*/*FLNC*) explained 3% of PC1 variance and 3.2% of IQ-adjusted PC1 variance in our largest GWAS dataset (UK-RD), and 1.3% of PC1 variance and 1.5% of IQ-adjusted PC1 variance in the next largest dataset (CLDRC-RD). The estimated effect sizes in the smaller datasets were ≤ 0.2%. Estimated effect sizes for rs5995177 (*RBFOX2*) were more consistent across datasets. This SNP explained 1.2% of PC1 and IQ-adjusted PC1 variance in UK-RD, and 1.8% of PC1 variance and 1.2% of IQ-adjusted PC1 variance in CLDRC-RD, while estimated effect sizes in the smaller datasets were between 0.6% and 1.6% of variance.

Both rs59197085 and rs5995177 showed broad profiles of association across the measures that were used to construct PC1, as assessed from the PLINK multivariate loadings and corresponding QFAM univariate association p-values shown in Table 4a, b. These findings suggest pleiotropic effects of the two SNPs on reading and language.

Gene-based meta-analysis

The strongest gene-based associations inferred from the PC1 and IQ-adjusted PC1 meta-analyses are reported in Tables S2c, d. While no gene exceeded the appropriate genome-wide significance threshold for this analysis ($p < 2.8\times10^{-6}$), *CCDC136*, *FLNC* and *RBFOX2* were among the most significantly associated genes, with the latter approaching the significance threshold in the PC1 analysis ($p = 5x10^{-6}$). However, after conditioning on the most significant association signal within each gene, no other SNP within each of these genes showed significant evidence for having an independent residual effect, after correction for multiple testing (lowest association $p \sim 0.028$, data not shown). For this analysis the gene boundaries were defined in the same way as for gene-based analysis (see above).

Pathway-based meta-analysis

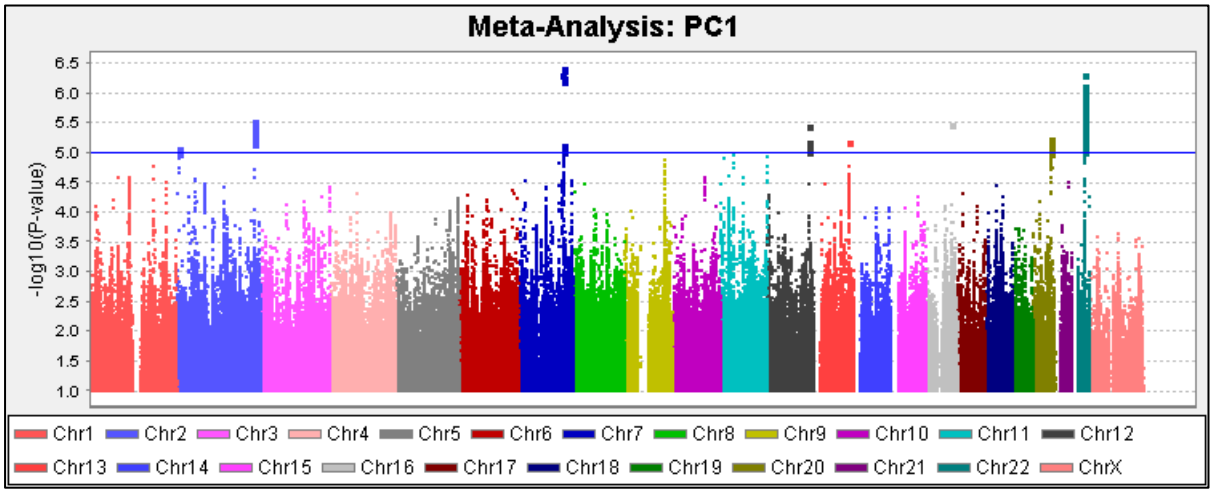
We assessed evidence for an excess of association signals from our GWASMA within the genes of three neurobiological pathways that are prominent in prior literature on reading and language: axon guidance, neuronal migration and steroid sex hormone biology (see *Discussion* for the relevant citations). None of the three tested gene sets were significantly associated with PC1 or IQ-adjusted PC1 (Table S2e, f), although the association between PC1 and the steroid-related pathway approached significance ($p = 0.051$).

Assessment of top associations within previous GWAS results

We assessed our most significant associations from PC1 meta-analyses within published and unpublished results from the previous GWAS study of the *BATS/ALSPAC* datasets, for which the reading and language measures were IQ-adjusted (Luciano et al. 2013). *FLNC* and *CCDC136* showed nominally significant associations in gene-based (VEGAS) analyses of reading-related traits in *BATS/ALSPAC* (*CCDC136* $p = 0.034$ for reading-spelling factor and $p = 0.003$ for word reading; *FLNC* $p = 0.009$ for word reading; see Table S3 of Luciano et al.

2013). The reading-spelling factor in the *BATS*/*ALSPAC* datasets was the most comparable trait to the IQ-adjusted PC1 score of the present study. As the study of Luciano et al. 2013 had used the HapMap2 reference dataset for genotype imputation, it was not possible to directly investigate the most highly-associated SNPs from the present study in the *BATS/ALSPAC* datasets. We therefore investigated association for two HapMap2 SNPs that were closest to our top hits on 7q32 and 22q12.3. rs3734972 (PC1 $p = 5.66x10^{-7}$, IQ-adjusted PC1 $p = 4.68x10^{-7}$; T/C, minor allele T, MAF ≈ 8%) lies ~10 kb away from rs59197085 on 7q32 and is in high LD with it (R^2 =0.89, see local association plot, Fig. S2c). rs3734972 showed a p-value of 0.032 with the IQ-adjusted reading-spelling factor in *BATS/ALSPAC*. The allelic trend was in the opposite direction to that observed in the *UK-RD/SLIC/CLDRC* datasets, with the T allele having a positive effect on the trait score in the *BATS/ALSPAC* cohorts. rs12158565 (PC1 $p = 7.57x10^{-7}$, IQ-adjusted PC1 $p = 4.65x10^{-5}$; C/G, minor allele G, MAF ≈ 13%) was the second most significant association in 22q12.3, mapping ~7 kb from the top SNP at this locus rs5995177, and in low LD with it (R^2 = 0.083), as are all the other suggestively associated SNPs in 22q12.3 (see local association plot, Fig. S2d). rs12158565 showed no evidence of association in *BATS/ALSPAC* (p = 0.81).

a)



b)

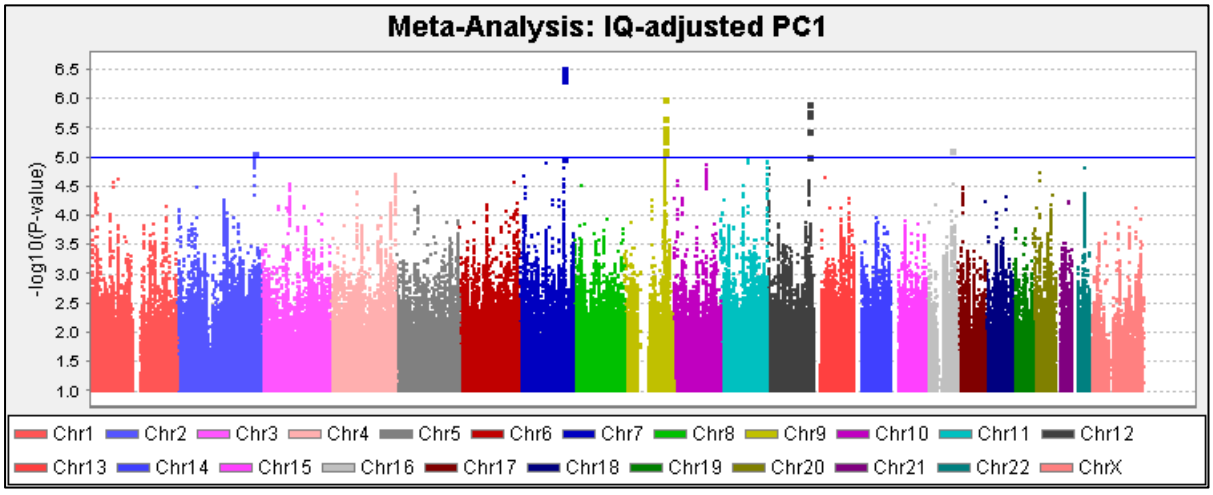


Figure 1. Manhattan plots of the **a)** PC1 and **b)** IQ-adjusted PC1 meta-analyses. The blue line represents the nominal suggestive significance threshold ($p = 1x10^{-5}$).

3a)

Chr	SNP[a]	Position (hg19)	Allele1	Allele2	Freq Allele1 (%)	P-value	Direction[b]	Gene (distance)[c]	Variant type
7	rs59197085	128460756	a	g	7.97	3.86×10^{-7}	----	FLNC(-9.726)|CCDC136(0)	intronic
7	rs58845495	128462847	t	c	92.03	4.09×10^{-7}	++++	FLNC(-7.635)|CCDC136(+0.664)	
7	7:128439695:I	128439695	i	r	7.94	4.99×10^{-7}	----	CCDC136(0)	intronic
22	rs5995177	36309553	a	g	8.05	5.01×10^{-7}	----	RBFOX2(0)	intronic
7	rs3734972	128470838	t	c	7.98	5.66×10^{-7}	----	FLNC(0)|CCDC136(+8.655)	exonic, synonymous
7	rs3800560	128461094	t	c	7.97	6.25×10^{-7}	----	FLNC(-9.388)|CCDC136(0)	intronic
22	rs12158565	36316843	c	g	87.23	7.57×10^{-7}	++++	RBFOX2(0)	intronic
22	rs5755979	36290707	t	c	12.77	9.05×10^{-7}	----	RBFOX2(0)	intronic
22	rs5750202	36339542	t	c	12.77	9.06×10^{-7}	----	RBFOX2(0)	intronic
22	rs5750203	36339998	a	t	87.23	9.72×10^{-7}	++++	RBFOX2(0)	intronic

3b)

Chr	SNP[a]	Position (hg19)	Allele1	Allele2	Freq Allele1 (%)	P-value	Direction[b]	Gene (distance)[c]	Variant type
7	rs59197085	128460756	a	g	7.97	3.01×10^{-7}	--+-	FLNC(-9.726)|CCDC136(0)	intronic
7	rs58845495	128462847	t	c	92.03	3.23×10^{-7}	++-+	FLNC(-7.635)|CCDC136(+0.664)	
7	rs3800560	128461094	t	c	7.97	3.95×10^{-7}	--+-	FLNC(-9.388)|CCDC136(0)	intronic
7	7:128439695:I	128439695	i	r	7.94	4.48×10^{-7}	--+-	CCDC136(0)	intronic
7	rs3734972	128470838	t	c	7.98	4.68×10^{-7}	--+-	FLNC(0)|CCDC136(+8.655)	exonic, synonymous

Table 3. Top association signals ($p < 1x10^{-6}$) in the **a)** PC1 and **b)** IQ-adjusted PC1 meta-analyses. [a] Single-base indels were not filtered out from the imputed polymorphisms since they were reliably called in the imputation reference (1000 Genomes, Phase I v3), and were tested for association as they could represent coding frameshift variants of biological interest. [b] The direction of effect of Allele1 is reported for datasets in the following order: CLDRC-RD, UK-RD, SLIC, CLDRC-ADHD. [c] Physical distance (kb) from closest genes (in a ±10kb range from each marker) is indicated, along with orientation based on the direction of transcription ("-" = upstream of 5'-UTR, "+" = downstream of 3'-UTR).

4a)

Trait[a]	CLDRC-RD	UK-RD	SLIC	CLDRC-ADHD
WRead	-0.66 (0.024)	-0.87 (5.3 x 10^{-5})	-0.29 (0.626)	-0.5 (0.427)
WSpell	-0.89 (3.8 x 10^{-3})	-0.75 (1.1 x 10^{-3})	0.08 (0.862)	-0.1 (0.871)
PD	-0.76 (7.9 x 10^{-3}), -0.50 (0.081) [b]	-0.86 (1.6 x 10^{-5})		-0.37 (0.549), 0.13 (0.854) [b]
PA	-0.65 (0.029)	-0.49 (0.018) [c]		0.35 (0.588)
OC	-0.64 (0.036)	-0.89 (3 x 10^{-6})		-0.04 (0.95)
NWR	-0.34 (0.269)		-0.57 (0.32)	-0.28 (0.686)
ELS			-0.25 (0.807)	
RLS			0.08 (0.821)	

4b)

Trait[a]	CLDRC-RD	UK-RD	SLIC	CLDRC-ADHD
WRead	-0.66 (0.027)	-0.81 (2 x 10^{-3})	-0.71 (0.116)	0.01 (0.98)
WSpell	-0.81 (6.9 x 10^{-3})	-0.82 (1.1 x 10^{-3})	-0.52 (0.262)	-0.33 (0.359)
PD	-0.65 (0.026), -0.79 (8.9 x 10^{-3}) [b]	-0.77 (1.8 x 10^{-3})		-0.46 (0.158), -0.37 (0.26) [b]
PA	-0.72 (0.023)	-0.72 (2.5 x 10^{-3}) [c]		-0.65 (0.046)
OC	-0.68 (0.026)	-0.57 (0.017)		-0.02 (0.968)
NWR	-0.04 (0.922)		-0.23 (0.674)	0.06 (0.876)
ELS			-0.82 (0.057)	
RLS			-0.61 (0.206)	

Table 4. Effect of the top association signals **a)** rs59197085 (7q32.1) and **b)** rs5995177 (22q12.3) on the single reading and language traits used in constructing PC1. These were computed for each trait as PLINK Multivariate MQFAM loadings and PLINK univariate QFAM association p-values (in brackets) and refer to the minor alleles (A for both SNPs).

[a] Legend: WRead = word reading; WSpell = word spelling; PD = phonological decoding; PA = phoneme awareness; OC = orthographic coding; NWR = nonword repetition; ELS/RLS = expressive/receptive language score. [b] Loading on nonword reading and phonological choice (respectively). [c] Although PA had been excluded from the PCA in UK-RD (due to the low number of measures available, see Chapter 2), it was tested in this case to have a term of comparison to the other datasets.

Discussion

The present study aimed to identify pleiotropic variants having effects on reading and language abilities by analyzing continuous traits in multiple datasets. Our study is complementary to two recently published GWAS: one using a similar approach in general population samples (Luciano et al. 2013), and another contrasting a relatively small number of categorically defined RD-SLI comorbid cases and unaffected controls (Eicher et al., 2013).

Our study is novel and distinct for several reasons. First, we analysed continuous variation in reading and language skills while also having an enrichment of participants with low abilities

(i.e. through analyzing poor performing probands together with their siblings), and without applying a dichotomous classification into cases and controls that necessarily involves arbitrary thresholding. Our design was therefore suited to detect genetic effects on susceptibility to RD and SLI that also act across the entire distribution of reading and language skills. Second, we specifically focused on shared neurobiological mechanisms underlying language and reading, by analyzing the first principal component of all of the reading- and language-related measures available in each dataset, followed by investigating the cross-phenotypic effects of the resulting top GWAS hits through univariate association analysis using each individual measure. We additionally followed this with a confirmatory analysis focused only on word reading and spelling, since these measures provided the closest matching possibility across our datasets. The first principal component (PC1) of all available measures extracted a large proportion of shared trait variance across the domains of reading and language, and was highly correlated with the component derived from only reading and spelling ($PC1_{read}$), as demonstrated in Chapter 2. Third, we performed GWAS both before and after IQ-adjustment of PC1. This was done in order to identify both genetic variants having effects broadly across reading, language and general cognitive abilities, and variants having effects on reading and language but independently of general cognitive ability. This approach also facilitated a comparison of our top results with those from datasets investigated in Luciano et al. (2013).

We checked within our GWASMA results 18 specific SNPs that had been highlighted to show the most promising candidate associations by the authors of previous GWAS studies of reading and/or language (Meaburn et al., 2008; Roeske et al., 2011; Field et al., 2013; Luciano et al., 2013; Eicher et al., 2013). Seventeen of these SNPS showed no nominally significant association within our GWASMA results (data not shown). Only rs10485609 (Meaburn et al., 2008) showed a nominally significant association ($p = 0.013$ for PC1, $p = 0.015$ for IQ-adjusted PC1; allele A was associated with lower performance, which was a consistent allelic direction of effect with that reported by Meaburn et al. 2008), but this was not significant after multiple testing correction for eighteen tests.

Like the other recently published GWAS efforts in this field, our study did not find any individual associations that achieved genome-wide significance (threshold $P = 5x10^{-8}$). However, we did identify two novel, suggestive results of particular interest, on 7q32.1 and 22q12.3, with the most significant associations at rs59197085 and rs5995177 respectively. As shown in Table 4, both SNPs displayed a broad pattern of association across multiple reading

and language traits, consistent with effects on neurobiological processes shared between reading and language cognition. In the regression model these SNPs explained a notable proportion (up to 3.2%) of variance in PC1 and IQ-adjusted PC1 scores, particularly in the largest datasets (CLDRC-RD and UK-RD), although these effect sizes are likely to be overestimated since this is the first report of these associations (Ioannidis, 2008). Gene based-tests were consistent with the results of the SNP-based analysis for *FLNC*, *CCDC136*, and *RBFOX2,* and the gene-based P values were found to be largely or wholly reflective of the individual top associations within each of these genes.

rs5995177 is an intronic variant localized within *RBFOX2* (RNA-binding protein, fox-1 homolog 2, also known as *RBM9*), a protein that regulates alternative splicing and is active in neurons. *RBFOX2* is highly expressed in the fetal brain and has important roles in CNS development (Gehman et al., 2012). The homologous gene *RBFOX1* has been implicated in several neurodevelopmental disorders, including Rolandic Epilepsy (Lal et al., 2013) and Autism Spectrum Disorder (Voineagu et al., 2011), and is a downstream target of FOXP2, a transcription factor implicated in monogenic speech and language disorders (Ayub et al., 2013). The high comorbidity between Rolandic Epilepsy and RD and SLI (Clarke et al., 2007; Pal, 2011), and the presence of a FOXP2 binding site ~1 kb from rs5995177 (The ENCODE Project Consortium, 2012), further support a link of *RBFOX2* with reading and language abilities. Thus convergent evidence from multiple lines of research makes *RBFOX2* an intriguing candidate gene for future studies. There was no evidence of association of this locus with reading and language measures in the results of the population-based study of Luciano et al. (2013).

rs59197085 is located in *CCDC136* (coiled-coil domain containing 136, or *NAG6*) and ~10 kb upstream of *FLNC* (filamin C). This SNP, along with the nearby SNPs rs3800560, rs58845495 and rs3734972, forms roughly 10-kb haplotypes spanning the region between *CCDC136* and *FLNC* and partially overlapping these genes (see local association plot, Fig. S2c). *CCDC136* encodes a poorly characterized tumor suppressor which has been found to be down-regulated in gastric carcinoma (Zhang et al., 2004) and is highly expressed in the cerebellum and in the occipital cortex (Allen Human Brain Atlas, Hawrylycz et al., 2012; http://human.brain-map.org). Filamin C (or filamin gamma) is a structural protein that crosslinks actin filaments into orthogonal networks in the cortical cytoplasm and participates in cytoskeleton re-modelling, suggesting a possible role in cell motility and migration. Functions of *FLNC* have been demonstrated in muscle tissues, where mutations are

responsible for several forms of myopathies (Duff et al., 2011). However, its pattern of expression includes spinal cord, cerebellum, corpus callosum, basal ganglia and some localized areas in the frontal, temporal and occipital cortex (Allen Human Brain Atlas, Hawrylycz et al., 2012). Its homologue *FLNA* (filamin A) is involved in neuronal migration and is implicated in an X-linked dominant form of periventricular heterotopia, a neurological disorder that sometimes involves reading and spelling problems (Robertson, 2005).

Associations within the 7q32 region are particularly interesting in light of data from two previous independent studies that have each reported evidence for linkage between a microsatellite marker in this region (D7S530, located ~650 kb from our peaks of association) and RD status (Kaminen et al., 2003) or else nonword spelling and irregular word reading (Bates et al., 2007). There was also evidence of association, at the gene level, with reading and language measures for *FLNC*, and *CCDC136* in the *BATS/ALSPAC* datasets studied by Luciano et al. (2013). At the SNP level, one of our most significantly associated SNPs from GWASMA, rs3734972, also showed association with an IQ-adjusted reading-spelling score in the *BATS*/*ALSPAC* datasets. However, the allelic directions of effect on the traits in the present study and the study by Luciano et al. were opposite.

We sought to detect an excess of association signals within genes belonging to each of three candidate gene sets based on different biological functions: axon guidance, neuronal migration, and steroid hormone biology. Axon guidance and neuronal migration are functions linked to some of the previously identified candidate genes in RD and SLI; *ROBO1* (Hannula-Jouppi et al., 2005), *DCDC2* (Meng et al., 2005), *KIAA0319* (Peschansky et al., 2010), *DYX1C1* (Tammimies et al., 2013) and *FOXP2* (Vernes et al., 2011). A potential involvement of neuronal migration deficits in RD etiology represents a longstanding hypothesis of the field (see Galaburda & Cestnick, 2003). The steroid hypothesis was motivated by literature suggesting links between sex hormone biology, language performance, and the brain architecture that subserves reading and language (Good et al., 2001; Shapleske et al., 1999; Whitehouse et al., 2012; Lombardo et al., 2012); and by evidence of interaction between Estrogen Receptors and DYX1C1, both at the gene (Tammimies et al., 2012) and at the protein level (Massinen et al., 2009). None of the three gene sets showed a significant excess of association signals, although the steroid hormone biology set approached significance in this analysis.

In carrying out GWASMA studies of complex cognitive traits across multiple datasets collected by different research teams, an obvious limitation is that the specific trait

measurements that are available may be quite diverse. Even when tests are similar, and hypothesized to measure corresponding cognitive processes, they may still create a substantial source of heterogeneity for a meta-analysis effort. In the present study we sought to overcome this limitation by focusing on a principal component (PC1) capturing a majority of the shared variance between reading- and language-related traits. In spite of the phenotypic heterogeneity of our datasets, this measure can be considered comparable across datasets for a number of reasons, detailed in Chapter 2. Firstly, the loadings of the individual traits on PC1 scores were similar across the datasets. Secondly, dropping one or more traits from our PC1 computation did not substantially affect the resulting PC1 scores. Thirdly, the First Principal Component derived only from word reading and spelling ($PC1_{read}$) was strongly correlated with PC1. Word reading and spelling were the only two measures available in all of the datasets and provided the closest phenotype matching possible across datasets. Not surprisingly, given the high correlations between PC1 and $PC1_{read}$ in all datasets, the association meta-analysis using $PC1_{read}$ (*Supplementary Material S3*) produced results consistent with PC1-based meta-analysis. We therefore conclude that PC1 was a sufficiently well matched construct across datasets to support GWASMA, in which we nonetheless allowed for heterogeneity of effect sizes across datasets to avoid assuming a perfect matching. It is interesting that a single PC can capture comparable variation across a diverse range of reading and language traits and in the presence of heterogeneity of measurement across datasets. This indicates a robust unifying dimension to much of this variation, and supports a genetic approach framed around pleiotropy.

The use of a principal component can lead to some loss of information, both in terms of detecting trait-specific genetic effects, and of reducing the sample size (since individuals with one or more missing trait values were excluded from the analysis). However, as we aimed to identify shared genetic effects on reading and language, the use of PC1 scores, followed by investigating cross-phenotypic associations of the top SNPs at the level of individual traits, was an appropriate approach to analyzing these multivariate datasets. There is now a need for a larger international meta-analysis effort that incorporates further datasets. This would improve the power to detect pleiotropic variants affecting reading and language.

Acknowledgements

The authors declare no conflicts of interest. This work was supported by the Max Planck Society, the University of St Andrews, the EU (Neurodys, 018696), and the US National Institutes of Health (Grant ref: P50 HD027802). Genotyping at the Wellcome Trust Centre for Human Genetics was supported by the Wellcome Trust (090532/Z/09/Z) and a Medical Research Council Hub Grant (G0900747 91070). Silvia Paracchini is a Royal Society University Research Fellow. Dianne Newbury is an MRC Career Development Fellow and a Junior Research Fellow at St John's College, University of Oxford. We thank Dr Manuel Ferreira and Dr Phil H Lee for useful advice regarding the use of PLINK Multivariate and INRICH tools. Dr Margaret Wright and Prof Nicholas Martin are the principal investigators for the BATS IQ data collection and genotyping. Members of the SLI Consortium: *Wellcome Trust Centre for Human Genetics, Oxford:* D. F. Newbury, N. H. Simpson, R. Nudel, A. P. Monaco; *Max Planck Institute for Psycholinguistics, Nijmegen:* S. E. Fisher, C. Francks; *Newcomen Centre, Guy's Hospital, London:* G. Baird, V. Slonims, K Dworzynski; *Child and Adolescent Psychiatry Department and Medical Research Council Centre for Social, Developmental, and Genetic Psychiatry, Institute of Psychiatry, London:* P. F. Bolton; *Medical Research Council Centre for Social, Developmental, and Genetic Psychiatry Institute of Psychiatry, London:* E. Simonoff; *Department of Reproductive and Developmental Sciences, University of Edinburgh:* A. O'Hare; *Molecular Medicine Centre, University of Edinburgh:* J. Seckl; *Department of Speech and Language Therapy, Royal Hospital for Sick Children, Edinburgh*: H. Cowie; *Speech and Hearing Sciences, Queen Margaret University College:* A. Clark and J. Watson; *Department of Educational and Professional Studies, University of Strathclyde:* W. Cohen; *Department of Child Health, the University of Aberdeen*: A. Everitt, E. R. Hennessy, D. Shaw, P. J. Helms; *Audiology and Deafness, School of Psychological Sciences,University of Manchester:* Z. Simkin, G. Conti-Ramsden; *Department of Experimental Psychology, University of Oxford:* D. V. M. Bishop; *Biostatistics Department, Institute of Psychiatry, London:* A. Pickles. ALSPAC: We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. The UK Medical Research Council and the Wellcome Trust (Grant ref: 092731) and the University of Bristol provide core support for ALSPAC.

Supplementary Material

- *S1: Supplementary Methods.* Genotype Calls and QC protocols; Statistical analyses, commands and parameters.
- *S2*: *Supplementary Results, PC1 meta-analysis.* QQ plots and association plots of the top association signals from analysis based on PC1 and IQ-adjusted PC1. SNP-, gene- and pathway-based meta-analysis results. Contribution of each dataset to the strength of the association in the PC1 and IQ-adjusted PC1 meta-analysis, for the top association signals.
- *S3: Supplementary Results, PC1$_{read}$ meta-analysis.* Analyses based on PC1$_{read}$ and IQ-adjusted PC1$_{read}$. Manhattan Plots, QQ plots and top associations.

References

Abecasis, G.R., Cardon, L.R. & Cookson, W.O.C. (2000) A General Test of Association for Quantitative Traits in Nuclear Families. *Am J Hum Genet ,* **66,** 279–292.

Abraham, J., Maranian, M., Spiteri, I., Russell, R., Ingle, S., Luccarini, C., Earl, H., Pharoah, P., Dunning, A. & Caldas, C. (2012) Saliva samples are a viable alternative to blood samples as a source of DNA for high throughput genotyping. *BMC Med Genomics*, **5**, 19.

Anthoni, H., Zucchelli, M., Matsson, H., Muller-Myhsok, B., Fransson, I., Schumacher, J., Massinen, S., Onkamo, P., Warnke, A., Griesemann, H., Hoffmann, P., Nopola-Hemmi, J., Lyytinen, H., Schulte-Korne, G., Kere, J., Nothen, M.M. & Peyrard-Janvid, M. (2007) A locus on 2p12 containing the co-regulated MRPL19 and C2ORF3 genes is associated to dyslexia. *Hum Mol Genet*, **16**, 667-677.

Ayub, Q., Yngvadottir, B., Chen, Y., Xue, Y., Hu, M., Vernes, Sonja C., Fisher, Simon E. & Tyler-Smith, C. (2013) FOXP2 Targets Show Evidence of Positive Selection in European Populations. *Am J Hum Genet,* **92,** 696-706.

Barry, J.G., Yasin, I. & Bishop, D.V. (2007) Heritable risk factors associated with language impairments. *Genes Brain Behav,* **6,** 66-76.

Bates, T.C., Luciano, M., Castles, A., Coltheart, M., Wright, M.J. & Martin, N.G. (2007) Replication of reported linkages for dyslexia and spelling and suggestive evidence for novel regions on chromosomes 4 and 17. *Eur J Hum Genet,* **15,** 194-203.

Bates, T.C., Luciano, M., Medland, S.E., Montgomery, G.W., Wright, M.J. & Martin, N.G. (2011) Genetic variance in a component of the language acquisition device: ROBO1 polymorphisms associated with phonological buffer deficits. *Behav Genet,* **41,** 50-57.

Becker, J., Czamara, D., Scerri, T.S., Ramus, F., Csepe, V., Talcott, J.B., Stein, J., Morris, A., Ludwig, K.U., Hoffmann, P., Honbolygo, F., Toth, D., Fauchereau, F., Bogliotti, C., Iannuzzi, S., Chaix, Y., Valdois, S., Billard, C., George, F., Soares-Boucaud, I., Gerard, C.L., van der Mark, S., Schulz, E., Vaessen, A., Maurer, U., Lohvansuu, K., Lyytinen, H., Zucchelli, M., Brandeis, D., Blomert, L., Leppanen, P.H., Bruder, J., Monaco, A.P., Muller-Myhsok, B., Kere, J., Landerl, K., Nothen, M.M., Schulte-Korne, G., Paracchini, S., Peyrard-Janvid, M. & Schumacher, J. (2013) Genetic analysis of dyslexia candidate genes in the European cross-linguistic NeuroDys cohort. *Eur J Hum Genet*, **22**, 675-680.

Bishop, D.V. (1994) Is specific language impairment a valid diagnostic category? Genetic and psycholinguistic evidence. *Phi Trans R Soc B,* **346,** 105-111.

Bishop, D.V.M. & Snowling, M.J. (2004) Developmental Dyslexia and Specific Language Impairment: Same or Different? *Psychol Bull*, **130**, 858-886.

Boyd, A., Golding, J., Macleod, J., Lawlor, D. A., Fraser, A., Henderson, J., et al. (2013). Cohort Profile: The 'Children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol*, **42**(1), 111-127.

Brandler, W.M., Morris, A.P., Evans, D.M., Scerri, T.S., Kemp, J.P., Timpson, N.J., St Pourcain, B., Smith, G.D., Ring, S.M., Stein, J., Monaco, A.P., Talcott, J.B., Fisher, S.E., Webber, C. & Paracchini, S. (2013) Common variants in left/right asymmetry genes and pathways are associated with relative hand skill. *PLoS Genet,* **9,** e1003751.

Carrion-Castillo, A., Franke, B. & Fisher, S.E. (2013) Molecular Genetics of Dyslexia: An Overview. *Dyslexia,* **19,** 214-240.

Ceroni, F., Simpson, N.H., Francks, C., Baird, G., Conti-Ramsden, G., Clark, A., Bolton, P.F., Hennessy, E.R., Donnelly, P., Bentley, D.R., Martin, H., Parr, J., Pagnamenta, A.T., Maestrini, E., Bacchelli, E., Fisher, S.E. & Newbury, D.F. (2014) Homozygous microdeletion of exon 5 in ZNF277 in a girl with specific language impairment. *Eur J Hum Genet*, **22**, 1165–1171.

Clarke, T., Strug, L.J., Murphy, P.L., Bali, B., Carvalho, J., Foster, S., Tremont, G., Gagnon, B.R., Dorta, N. & Pal, D.K. (2007) High Risk of Reading Disability and Speech Sound Disorder in Rolandic Epilepsy Families: Case–Control Study. *Epilepsia*, **48**, 2258-2265.

Cope, N., Harold, D., Hill, G., Moskvina, V., Stevenson, J., Holmans, P., Owen, M.J., O'Donovan, M.C. & Williams, J. (2005) Strong Evidence That KIAA0319 on Chromosome 6p Is a Susceptibility Gene for Developmental Dyslexia. *Am J Hum Genet ,* **76,** 581-591.

DeFries, J.C., Filipek, P.A., Fulker, D.W., Olson, R.K., Pennington, B.F., Smith, S.D. & Wise, B.W. (1997) Colorado Learning Disabilities Research Center. *Learn Disabil Multidisciplinary J,* **8,** 7-19.

Duff, R.M., Tay, V., Hackman, P., Ravenscroft, G., McLean, C., Kennedy, P., Steinbach, A., Schoffler, W., van der Ven, P.F., Furst, D.O., Song, J., Djinovic-Carugo, K., Penttila, S., Raheem, O., Reardon, K., Malandrini, A., Gambelli, S., Villanova, M., Nowak, K.J., Williams, D.R., Landers, J.E., Brown, R.H., Jr., Udd, B. & Laing, N.G. (2011) Mutations in the N-terminal actin-binding domain of filamin C cause a distal myopathy. *Am J Hum Genet,* **88,** 729-740.

Eicher, J.D., Powers, N.R., Miller, L.L., Akshoomoff, N., Amaral, D.G., Bloss, C.S., Libiger, O., Schork, N.J., Darst, B.F., Casey, B.J., Chang, L., Ernst, T., Frazier, J., Kaufmann, W.E., Keating, B., Kenet, T., Kennedy, D., Mostofsky, S., Murray, S.S., Sowell, E.R., Bartsch, H., Kuperman, J.M., Brown, T.T., Hagler, D.J., Jr., Dale, A.M., Jernigan, T.L., Pourcain, B.S., Smith, G.D., Ring, S.M. & Gruen, J.R. (2013) Genome-Wide Association Study of Shared Components of Reading Disability and Language Impairment. *Genes Brain Behav*, **8,** 792-801.

Falcaro, M., Pickles, A., Newbury, D.F., Addis, L., Banfield, E., Fisher, S.E., Monaco, A.P., Simkin, Z. & Conti-Ramsden, G. (2008) Genetic and phenotypic effects of phonological short-term memory and grammatical morphology in specific language impairment. *Genes Brain Behav*, **7,** 393-402.

Ferreira, M.A. & Purcell, S.M. (2009) A multivariate test of association. *Bioinformatics*, **25**, 132-133.

Fulker, D.W., Cherny, S.S., Sham, P.C. & Hewitt, J.K. (1999) Combined Linkage and Association Sib-Pair Analysis for Quantitative Traits. *Am J Hum Genet ,* **64,** 259-267.

Field, A. (2005) *Discovering statistics using SPSS*, SAGE Publications Ltd, London.

Field, L.L., Shumansky, K., Ryan, J., Truong, D., Swiergala, E. & Kaplan, B.J. (2013) Dense-map genome scan for dyslexia supports loci at 4q13, 16p12, 17q22; suggests novel locus at 7q36. *Genes Brain Behav*, **12**, 56-69.

Fisher, S.E. & DeFries, J.C. (2002) Developmental dyslexia: genetic dissection of a complex cognitive trait. *Nat Rev Neurosci,* **3,** 767-780.

Fisher, S.E. & Scharff, C. (2009) FOXP2 as a molecular window into speech and language. *Trends Genet,* **25,** 166-177.

Francks, C., Paracchini, S., Smith, S.D., Richardson, A.J., Scerri, T.S., Cardon, L.R., Marlow, A.J., MacPhie, I.L., Walter, J., Pennington, B.F., Fisher, S.E., Olson, R.K., DeFries, J.C., Stein, J.F. & Monaco, A.P. (2004) A 77-Kilobase Region of Chromosome 6p22.2 Is Associated with Dyslexia in Families From the United Kingdom and From the United States. *Am J Hum Genet,* **75,** 1046-1058.

Gehman, L.T., Meera, P., Stoilov, P., Shiue, L., O'Brien, J.E., Meisler, M.H., Ares, M., Otis, T.S. & Black, D.L. (2012) The splicing regulator Rbfox2 is required for both cerebellar development and mature motor function. *Genes Dev,* **26,** 445-460.

Hannula-Jouppi, K., Kaminen-Ahola, N., Taipale, M., Eklund, R., Nopola-Hemmi, J., Kääriäinen, H. & Kere, J. (2005) The axon guidance receptor gene ROBO1 is a candidate gene for developmental dyslexia. *PLoS Genet,* **1**(4): e50.

Harlaar, N., Hayiou-Thomas, M.E., Dale, P.S. & Plomin, R. (2008) Why Do Preschool Language Abilities Correlate With Later Reading? A Twin Study. *J Speech Lang Hear Res*, **51**, 688-705.

Hawrylycz, M.J., Lein, E.S., Guillozet-Bongaarts, A.L., Shen, E.H., Ng, L., Miller, J.A., van de Lagemaat, L.N., Smith, K.A., Ebbert, A., Riley, Z.L., Abajian, C., Beckmann, C.F., Bernard, A., Bertagnolli, D., Boe, A.F., Cartagena, P.M., Chakravarty, M.M., Chapin, M., Chong, J., Dalley, R.A., Daly, B.D., Dang, C., Datta, S., Dee, N., Dolbeare, T.A., Faber, V., Feng, D., Fowler, D.R., Goldy, J., Gregor, B.W., Haradon, Z., Haynor, D.R., Hohmann, J.G., Horvath, S., Howard, R.E., Jeromin, A., Jochim, J.M., Kinnunen, M., Lau, C., Lazarz, E.T., Lee, C., Lemon, T.A., Li, L., Li, Y., Morris, J.A., Overly, C.C., Parker, P.D., Parry, S.E., Reding, M., Royall, J.J., Schulkin, J., Sequeira, P.A., Slaughterbeck, C.R., Smith, S.C., Sodt, A.J., Sunkin, S.M., Swanson, B.E., Vawter, M.P., Williams, D., Wohnoutka, P., Zielke, H.R., Geschwind, D.H., Hof, P.R., Smith, S.M., Koch, C., Grant, S.G.N. & Jones, A.R. (2012) An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*, **489**, 391-399.

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet*, **44**, 955-959.

Ioannidis, J.P.A. (2008) Why Most Discovered True Associations Are Inflated. *Epidemiology*, **19**, 640-648.

Johnson, A.D., Handsaker, R.E., Pulit, S.L., Nizzari, M.M., O'Donnell, C.J. & de Bakker, P.I.W. (2008) SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics,* **24,** 2938-2939.

Kaminen, N., Hannula-Jouppi, K., Kestilä, M., Lahermo, P., Muller, K., Kaaranen, M., Myllyluoma, B., Voutilainen, A., Lyytinen, H., Nopola-Hemmi, J. & Kere, J. (2003) A genome scan for developmental dyslexia confirms linkage to chromosome 2p11 and suggests a new locus on 7q32. *J Med Genet,* **40,** 340-345.

Lal, D., Reinthaler, E.M., Altmüller, J., Toliat, M.R., Thiele, H., Nürnberg, P., Lerche, H., Hahn, A., Møller, R.S., Muhle, H., Sander, T., Zimprich, F. & Neubauer, B.A. (2013) RBFOX1 and RBFOX3 Mutations in Rolandic Epilepsy. *PLoS One,* **8**(9): e73323.

Lee, P.H., O'Dushlaine, C., Thomas, B. & Purcell, S.M. (2012) INRICH: interval-based enrichment analysis for genome-wide association studies. *Bioinformatics,* **28,** 1797-1799.

Liu, J.Z., McRae, A.F., Nyholt, D.R., Medland, S.E., Wray, N.R., Brown, K.M., Hayward, N.K., Montgomery, G.W., Visscher, P.M., Martin, N.G. & Macgregor, S. (2010) A versatile gene-based test for genome-wide association studies. *Am J Hum Genet,* **87,** 139-145.

Logan, J., Petrill, S., Flax, J., Justice, L., Hou, L., Bassett, A., Tallal, P., Brzustowicz, L. & Bartlett, C. (2011) Genetic Covariation Underlying Reading, Language and Related Measures in a Sample Selected for Specific Language Impairment. *Behav Genet*, **41**, 651-659.

Luciano, M., Evans, D.M., Hansell, N.K., Medland, S.E., Montgomery, G.W., Martin, N.G., Wright, M.J. & Bates, T.C. (2013) A genome-wide association study for reading and language abilities in two population cohorts. *Genes Brain Behav,* **12,** 645-652.

Luciano, M., Lind, P.A., Duffy, D.L., Castles, A., Wright, M.J., Montgomery, G.W., Martin, N.G. & Bates, T.C. (2007) A haplotype spanning KIAA0319 and TTRAP is associated with normal variation in reading and spelling ability. *Biol Psychiatry,* **62,** 811-817.

McArthur, G.M., Hogben, J.H., Edwards, V.T., Heath, S.M. & Mengler, E.D. (2000) On the "Specifics" of Specific Reading Disability and Specific Language Impairment. *J Child Psychol Psychiatry,* **41,** 869-874.

Meaburn, E.L., Harlaar, N., Craig, I.W., Schalkwyk, L.C. & Plomin, R. (2008) Quantitative trait locus association scan of early reading disability and ability using pooled DNA and 100K SNP microarrays in a sample of 5760 children. *Mol Psychiatry*, **13**, 729-740.

Meng, H., Smith, S.D., Hager, K., Held, M., Liu, J., Olson, R.K., Pennington, B.F., DeFries, J.C., Gelernter, J., O'Reilly-Pol, T., Somlo, S., Skudlarski, P., Shaywitz, S.E., Shaywitz, B.A., Marchione, K., Wang, Y., Paramasivam, M., LoTurco, J.J., Page, G.P. & Gruen, J.R. (2005) DCDC2 is associated with reading disability and modulates neuronal development in the brain. *Proc Natl Acad Sci U S A,* **102,** 17053-17058.

Newbury, D.F. & Monaco, A.P. (2010) Genetic advances in the study of speech and language disorders. *Neuron,* **68,** 309-320.

Newbury, D.F., Paracchini, S., Scerri, T.S., Winchester, L., Addis, L., Richardson, A.J., Walter, J., Stein, J.F., Talcott, J.B. & Monaco, A.P. (2011) Investigation of dyslexia and SLI risk variants in reading- and language-impaired subjects. *Behav Genet,* **41,** 90-104.

Newbury, D.F., Winchester, L., Addis, L., Paracchini, S., Buckingham, L.L., Clark, A., Cohen, W., Cowie, H., Dworzynski, K., Everitt, A., Goodyer, I.M., Hennessy, E., Kindley, A.D., Miller, L.L., Nasir, J., O'Hare, A., Shaw, D., Simkin, Z., Simonoff, E., Slonims, V., Watson, J., Ragoussis, J., Fisher, S.E., Seckl, J.R., Helms, P.J., Bolton, P.F., Pickles, A., Conti-Ramsden, G., Baird, G., Bishop, D.V. & Monaco, A.P. (2009) CMIP and ATP2C2 modulate phonological short-term memory in language impairment. *Am J Hum Genet,* **85,** 264-272.

Nudel, R., Simpson, N.H., Baird, G., O'Hare, A., Conti-Ramsden, G., Bolton, P.F., Hennessy, E.R., the, S.L.I.C., Ring, S.M., Smith, G.D., Francks, C., Paracchini, S., Monaco, A.P., Fisher, S.E. & Newbury, D.F. (2014) Genome-wide association analyses of child genotype effects and parent-of-origin effects in specific language impairment. *Genes Brain Behav*, **13**, 418-429.

Pal, D.K. (2011) Epilepsy and neurodevelopmental disorders of language. *Curr Opin Neurol*, **24**, 126-131.

Paracchini, S., Steer, C.D., Buckingham, L.L., Morris, A.P., Ring, S., Scerri, T., Stein, J., Pembrey, M.E., Ragoussis, J., Golding, J. & Monaco, A.P. (2008) Association of the KIAA0319 Dyslexia Susceptibility Gene With Reading Skills in the General Population. *Am J Psychiatry,* **165,** 1576-1584.

Paracchini, S. (2011) Dissection of genetic associations with language-related traits in population-based cohorts. *J Neurodev Disord,* **3,** 365-373.

Paracchini, S., Ang, Q.W., Stanley, F.J., Monaco, A.P., Pennell, C.E. & Whitehouse, A.J. (2011) Analysis of dyslexia candidate genes in the Raine cohort representing the general Australian population. *Genes Brain Behav,* **10,** 158-165.

Pennington, B.F. (2006) From single to multiple deficit models of developmental disorders. *Cognition,* **101,** 385-413.

Pennington, B.F. & Bishop, D.V. (2009) Relations among speech, language, and reading disorders. *Annu Rev Psychol,* **60,** 283-306.

Peterson, R.L. & Pennington, B.F. (2012) Developmental dyslexia. *Lancet,* **379,** 1997-2007.

Poelmans, G., Buitelaar, J.K., Pauls, D.L. & Franke, B. (2011) A theoretical molecular network for dyslexia: integrating available genetic findings. *Mol Psychiatry,* **16,** 365-382.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. & Sham, P.C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet,* **81,** 559-575.

R Core Team (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Robertson, S.P. (2005) Filamin A: phenotypic diversity. *Curr Opin Genet Dev,* **15,** 301-307.

Roeske, D., Ludwig, K.U., Neuhoff, N., Becker, J., Bartling, J., Bruder, J., Brockschmidt, F.F., Warnke, A., Remschmidt, H., Hoffmann, P., Muller-Myhsok, B., Nothen, M.M. & Schulte-Korne, G. (2011) First genome-wide association scan on neurophysiological endophenotypes points to trans-regulation effects on SLC2A3 in dyslexic children. *Mol Psychiatry*, **16**, 97-107.

Scerri, T.S., Brandler, W.M., Paracchini, S., Morris, A.P., Ring, S.M., Richardson, A.J., Talcott, J.B., Stein, J. & Monaco, A.P. (2011a) PCSK6 is associated with handedness in individuals with dyslexia. *Hum Mol Genet,* **20,** 608-614.

Scerri, T.S., Morris, A.P., Buckingham, L.-L., Newbury, D.F., Miller, L.L., Monaco, A.P., Bishop, D.V.M. & Paracchini, S. (2011b) DCDC2, KIAA0319 and CMIP Are Associated with Reading-Related Traits. *Biol Psychiatry,* **70,** 237-245.

Schonell, F.J. & Schonell, P.E. (1960) *Diagnostic and Attainment Testing,* Oliver & Boyd, Edinburgh.

Shaywitz, S.E., Shaywitz, B.A., Fletcher, J.M.à. & Escobar, M.D. (1990) Prevalence of reading disability in boys and girls. Results of the Connecticut Longitudinal Study. *JAMA,* **264,** 998–1002.

Simpson, N.H., Addis, L., Brandler, W.M., Slonims, V., Clark, A., Watson, J., Scerri, T.S., Hennessy, E.R., Bolton, P.F., Conti-Ramsden, G., Fairfax, B.P., Knight, J.C., Stein, J., Talcott, J.B., O'Hare, A., Baird, G., Paracchini, S., Fisher, S.E., Newbury, D.F. & Consortium, S.L.I. (2014) Increased prevalence of sex chromosome aneuploidies in specific language impairment and dyslexia. *Dev Med Child Neurol*, **4**, 346-353.

Snowling, M., Bishop, D.V. & Stothard, S.E. (2000) Is preschool language impairment a risk factor for dyslexia in adolescence? *J Child Psychol Psychiatry. ,* **41,** 587–600.

Taipale, M., Kaminen, N., Nopola-Hemmi, J., Haltia, T., Myllyluoma, B., Lyytinen, H., Muller, K., Kaaranen, M., Lindsberg, P.J., Hannula-Jouppi, K. & Kere, J. (2003) A candidate gene for developmental dyslexia encodes a nuclear tetratricopeptide repeat domain protein dynamically regulated in brain. *Proc Natl Acad Sci USA,* **100,** 11553-11558.

The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature,* **491,** 56-65.

The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.

The International HapMap 3 Consortium. (2010) Integrating common and rare genetic variation in diverse human populations. *Nature,* **467,** 52-58.

The SLI Consortium (2002) A Genome wide Scan Identifies Two Novel Loci Involved in Specific Language

Impairment. *Am J Hum Genet ,* **70,** 384-398.

The SLI Consortium (2004) Highly significant linkage to the SLI1 locus in an expanded sample of individuals affected by specific language impairment. *Am J Hum Genet,* **74,** 1225-1238.

Tomblin, J.B., Records, N.L., Buckwalter, P., Zhang, X., Smith, E. & O'Brien, M. (1997) Prevalence of specific language impairment in kindergarten children. *J Speech Lang Hear Res,* **40,** 1245-1260.

Vernes, S.C., Newbury, D.F., Abrahams, B.S., Winchester, L., Nicod, J., Groszer, M., Alarcon, M., Oliver, P.L., Davies, K.E., Geschwind, D.H., Monaco, A.P. & Fisher, S.E. (2008) A functional genetic link between distinct developmental language disorders. *N Engl J Med,* **359,** 2337-2345.

Vernes, S.C., Oliver, P.L., Spiteri, E., Lockstone, H.E., Puliyadi, R., Taylor, J.M., Ho, J., Mombereau, C., Brewer, A., Lowy, E., Nicod, J., Groszer, M., Baban, D., Sahgal, N., Cazier, J.-B., Ragoussis, J., Davies, K.E., Geschwind, D.H. & Fisher, S.E. (2011) Foxp2 Regulates Gene Networks Implicated in Neurite Outgrowth in the Developing Brain. *PLoS Genet,* **7,** e1002145.

Voineagu, I., Wang, X., Johnston, P., Lowe, J.K., Tian, Y., Horvath, S., Mill, J., Cantor, R.M., Blencowe, B.J. & Geschwind, D.H. (2011) Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature,* **474,** 380-384.

Whitehouse, A.J., Bishop, D.V., Ang, Q.W., Pennell, C.E. & Fisher, S.E. (2011) CNTNAP2 variants affect early language development in the general population. *Genes Brain Behav,* **10,** 451-456.

Willcutt, E.G., Betjemann, R.S., McGrath, L.M., Chhabildas, N.A., Olson, R.K., DeFries, J.C. & Pennington, B.F. (2010) Etiology and neuropsychology of comorbidity between RD and ADHD: the case for multiple-deficit models. *Cortex,* **46,** 1345-1361.

Willcutt, E.G., Pennington, B.F., Olson, R.K., Chhabildas, N. & Hulslander, J. (2005) Neuropsychological analyses of comorbidity between reading disability and attention deficit hyperactivity disorder: in search of the common deficit. *Dev Neuropsychol,* **27,** 35-78.

Willer, C.J., Li, Y. & Abecasis, G.R. (2010) METAL: fast and efficient meta-analysis of genome wide association scans. *Bioinformatics,* **26,** 2190-2191.

Wright, M., De Geus, E., Ando, J., Luciano, M., Posthuma, D., Ono, Y., Hansell, N., Van Baal, C., Hiraishi, K., Hasegawa, T., Smith, G., Geffen, G., Geffen, L., Kanba, S., Miyake, A., Martin, N. & Boomsma, D. (2001) Genetics of Cognition: Outline of a Collaborative Twin Study. *Twin Res,* **4,** 48-56.

Zhang, X.M., Sheng, S.R., Wang, X.Y., Bin, L.H., Wang, J.R. & Li, G.Y. (2004) Expression of tumor related gene NAG6 in gastric cancer and restriction fragment length polymorphism analysis. *World J Gastroenterol* **10,** 1361-1364.

S1: Supplementary Methods

Genotype Calls and QC

In all the datasets both blood and saliva (including Oragene® kit) samples were collected, and within datasets some samples were genotyped from blood DNA while others were genotyped from saliva DNA. Comparable call rates and concordance rates between blood and saliva samples have been reported in the literature (Abraham et al., 2012).

Genotype Calls

UK-RD

For 200 subjects, referred to as *UK-RD_small* hereafter: Genotype calls were generated using Illumina® BeadStudio software from Illumina® HumanHap 550k v1 chip. Default settings as described in other studies (Scerri et al., 2011a) were used. 550,927 SNPs were finally called.

For an additional 818 samples, called *UK-RD_big* hereafter: Genotyping was implemented on the Illumina® Human OmniExpress (v12, manifest H, 730k) array. Genotype calls were made through Illumina® GenomeStudio software according to the following protocol:

1. all the SNPs mapped as "Y" (Y chromosome) and "0" (not mapped) were zeroed (i.e. set to missing);
2. samples with genotyping success rate < 95% were discarded;
3. SNPs with call frequency < 100% were re-clustered (i.e. their intensity data were re-plotted, in order to get better quality of the calls);
4. SNPs with call frequency < 99% were zeroed;
5. samples with updated genotyping success rate < 98% were excluded;
6. SNPs with *Cluster Sep* (i.e. measure of the cluster separation for a SNP, that ranges between 0 and 1 and indicates how well the intensity signals of the different genotypes are distinguishable) < 0.3 were zeroed.

Each passage was followed by an update in SNP/sample statistics, in order to improve the quality of the genotype calls and of the samples. At the end of this procedure, 716,044 SNPs were finally called.

SLIC

Samples were genotyped using the Illumina® Human Omni-Express (v12.1, manifest C) array, within the GenomeStudio software. Samples were randomized across plates, with probands and co-siblings being spread evenly across plates. Also sample types (blood vs saliva) have been randomized across plates and we checked for systematic differences between genotype and allele frequencies both between plates and samples types and all were non-significant. 47 samples were duplicated across plates (concordance rate 99.97%). SNPs and samples with a genotype success rate < 95% and/or heterozygosity rates ±2SD from the mean were removed, as were all SNPs with a Minor Allele Frequency (MAF) < 1%. SNPs with a *Gen Train* score (i.e. a number between 0 and 1 indicating how well the samples clustered for a specific locus) < 0.5 were removed. Since parents were also genotyped, SNPs and samples with an error rate ≥ 1%, as estimated by impossible inheritances within families, were removed (SLIC 2002; 2004; Newbury et al., 2009). A total of 630,167 SNPs were called.

CLDRC

Genotype calls were generated from Illumina® Human OmniExpress (v12, manifest H, 730k) array, using GenomeStudio software. The same protocol followed for *UK-RD_big* was used, finally resulting in genotypes for 683,242 total SNPs before quality control.

Genotype quality control (QC)

UK-RD

Pre-imputation QC. Since the two subsets of the *UK-RD* dataset had been genotyped on two different Illumina® platforms (as mentioned above), they were analyzed separately before imputation. However, in order to check for the absence of population stratification in the whole *UK-RD* dataset due to the different arrays used, the two subsets were temporarily merged and underwent a Multi-Dimensional Scaling (MDS) analysis of genome-wide SNP data (extracting the first 20 dimensions) on a subset of unrelated individuals (one subject per family selected from the whole dataset). This analysis revealed no effects of the factor mentioned above and received further support by the high genotype concordance rate (99.98%) of 34 duplicate samples genotyped in both subsets.

In *UK-RD_small*, 3 samples with low call rate (<98%) and 1 genome-wide homozygosity outlier (i.e. showing an extremely low homozygosity, which may suggest a bad quality of the DNA sample) were excluded. No Identity By Descent (IBD) sharing or sex inconsistencies between reported and genetically inferred information, nor MDS outliers (Figure S1a), were detected within this subset. We filtered out all the SNPs deviating from Hardy-Weinberg Equilibrium (HWE, p-val < $1x10^{-6}$, 12,631 SNPs) and with MAF < 1% (23,467 SNPs) in the whole subset (all unrelated individuals). 82,052 variants with call frequency < 99% were discarded.

In *UK-RD_big*, 7 samples with IBD sharing inconsistencies (half-siblings, unrelated samples showing cryptic relatedness or MZ twins); 8 sex mismatches (including 3 X chromosome abnormalities carriers) and 3 homozygosity outliers were excluded, along with 2 outliers in the MDS analysis on a subset of unrelated individuals (one subject per family, Figure S1b). All the samples had a call rate ≥ 98%. All the SNPs deviating from HWE (p-val < $1x10^{-6}$, 191 SNPs) and with MAF < 1% (77,342 SNPs) as calculated within the subset of unrelated individuals (one subject per family) were filtered out. No variants had call frequency < 99%.

A further IBD sharing check on the whole dataset revealed an inconsistency on one of the duplicated samples that had already been excluded in *UK-RD_big* but not in *UK-RD_small*, from which it was discarded.

Post-imputation QC. To ensure a high quality of imputation, imputed SNPs with r^2 (squared correlation between the allele count estimated for a given SNP by the imputation algorithm and the allele count that would be expected if the genotype of that SNP was observed without error) < 0.3 were discarded, and all the individual genotypes with quality score (estimated probability that an imputed genotype will match an experimental genotype) < 0.9 were set to missing. Then the two subsets were merged into the definitive *UK-RD* dataset (N = 959): 2,779 SNPs failed the HWE test (p-val < $5x10^{-6}$) and 1,980,500 had a MAF < 1%, in a subset of unrelated individuals (one subject per family); 1,704,412 SNPs were finally excluded due to call frequency < 95%, resulting in a final total of 6,190,549 SNPs analyzed in *UK-RD*. All the samples had a call rate ≥ 95%. MDS and IBD sharing analyses of imputed data confirmed the consistency with genotyped data (concordance rate before vs after imputation 99.96%, and 99.89% between the duplicated samples genotyped and imputed separately in the two subsets).

SLIC

Pre-imputation QC. Out of 548 subjects, 9 were excluded for sex chromosome abnormalities and 1 for X chromosome call rate <95% (Newbury, personal communication); 9 for genomic call-rate < 98%; 17 for IBD sharing typical of half-siblings (19-31%) when they had been reported as full siblings; 3 sex mismatches and 2 outliers on genome-wide homozygosity. An MDS analysis of genome-wide SNP data was run on a subset of unrelated individuals (one subject per family): 2 outliers were detected and excluded (Figure S1c), along with their 3 siblings (5 in total). We filtered out SNPs deviating from HWE (p-val < $1x10^{-6}$, 54 SNPs) and with MAF < 1% (1,718 SNPs) as calculated in the same subset of unrelated individuals (one subject per family), as well as (72,043) variants with call frequency < 99%.

Post-imputation QC. Imputed SNPs with $r^2 < 0.3$ were filtered out, and all the genotypes with quality score < 0.9 were set to missing. 2,096 SNPs deviated from HWE (p-val < $5x10^{-6}$) and 3,260,639 had a MAF < 1% in a subset of unrelated individuals (one individual per family); 1,766,376 SNPs were excluded for call frequency < 95%, leading to a final total of 6,240,842 SNPs analyzed. All the samples showed a call rate ≥ 95%. MDS and IBD sharing analyses of imputed data confirmed their consistency with directly genotyped data, as did the concordance rate between imputed and genotyped data (99.97%).

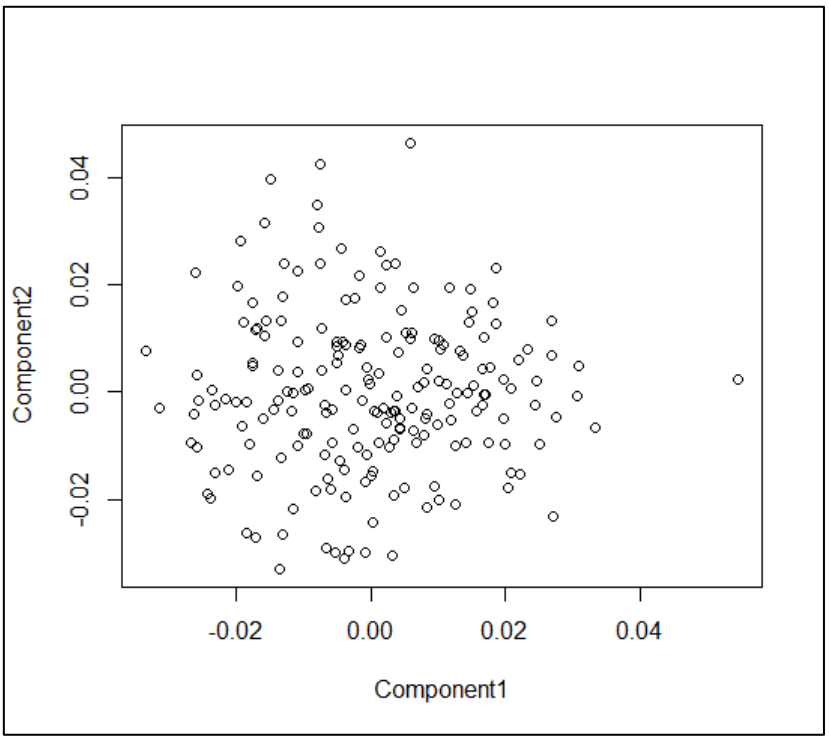
CLDRC

Pre-imputation QC. Since *CLDRC-RD* and *CLDRC-ADHD* belonged to the same dataset, we decided to treat them as a unique dataset in the genotype QC. Out of 749 initial *CLDRC* subjects, 11 samples with IBD sharing inconsistencies (half-siblings or unrelated samples showing cryptic relatedness), 3 sex mismatches, and 6 homozygosity outliers were discarded (all the samples had call rate ≥ 98% and there were no outliers in the MDS analysis on a subset of unrelated individuals, including one subject per family, Figure S1d). We filtered out all the SNPs deviating from HWE (p-val < $1x10^{-6}$, 57 SNPs) and with MAF < 1% (74,770 SNPs) in a subset of unrelated individuals (one subject per family). No variants had call frequency < 99%.

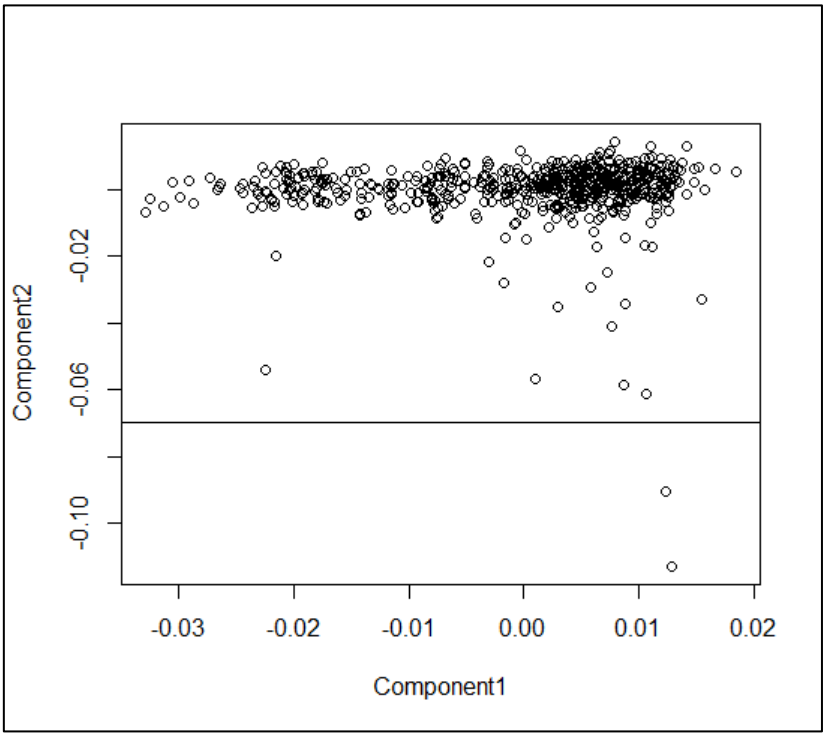
Post-imputation QC. Imputed SNPs with $r^2 < 0.3$ were filtered out, and all the genotypes with quality score < 0.9 were set to missing. 2,166 SNPs did not pass the HWE test (p-val < $5x10^{-6}$) and 3,640,742 had a MAF < 1%, in a subset of unrelated individuals (one subject per family); 1,729,493 SNPs were finally excluded for call frequency < 95%. All the samples

showed a call rate ≥ 95%. A total of 6,427,200 SNPs were examined in both the *CLDRC* datasets. MDS and IBD sharing analyses of imputed data confirmed the consistency with genotyped data (concordance rate 99.96% between genotyped and imputed data).

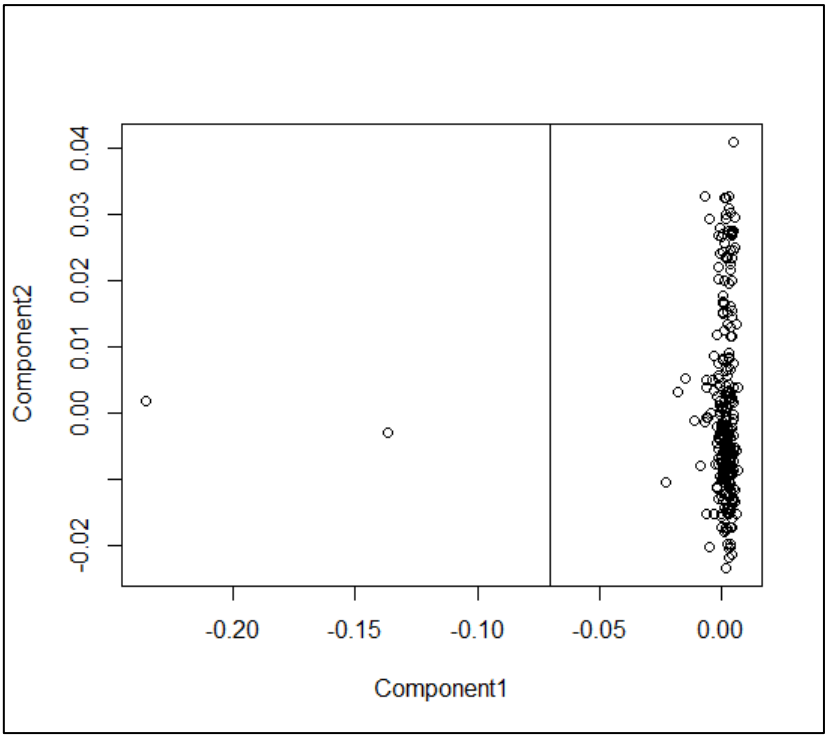
a)



b)



c)



d)

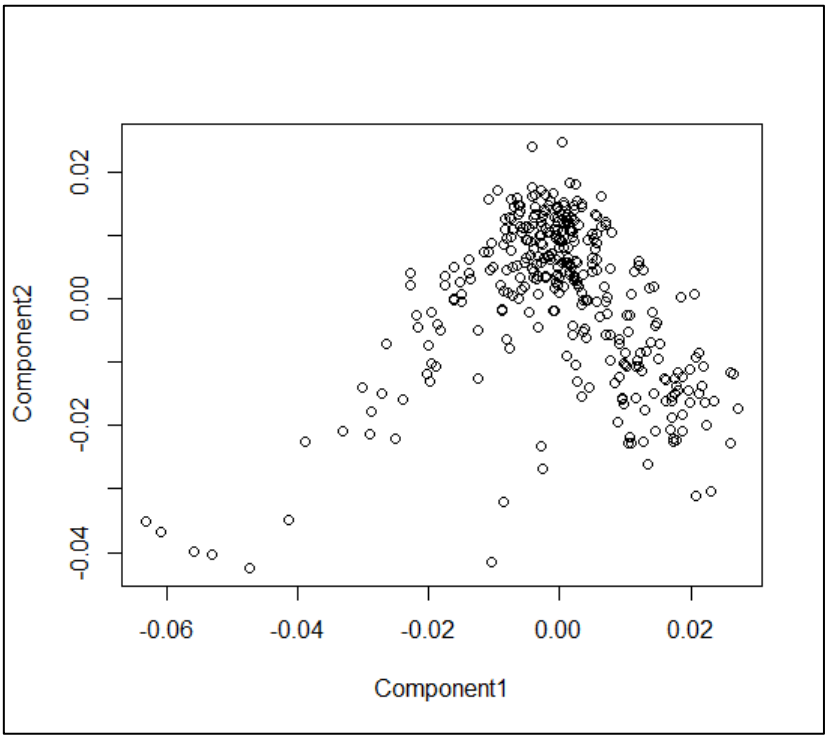


Figure S1. MDS analysis of **a)** *UK-RD_small* subset, **b)** *UK-RD_big* subset, **c)** SLIC and **d)** CLDRC samples on non-imputed data. The most evident outliers (i.e. showing at least one of the first 3 MDS components scores extracted out of the interval [-0.7; 0.7]), were excluded (along with their co-siblings, where present). For simplicity, only the first two MDS components are shown (Component 3 did not show any outlier in any of the datasets). Imputed samples passing QC in each dataset underwent a second MDS analysis which did not reveal any outlier (data not shown).

Statistical analyses: commands and parameters

PLINK QFAM (family-based association tests for quantitative traits) analysis

Sibling-based genome-wide association analysis of PC1 scores was conducted using PLINK v1.07 (Purcell et al., 2007) *--qfam-total* analysis, a permutation-based method correcting for subject relatedness. In this analysis the association between a quantitative trait and a SNP is tested by regressing the trait score on the SNP genotype in an additive model (as in the *--linear* analysis). However, to adjust for sample relatedness, a high number of permutations (i.e. label-swapping of phenotypes/genotypes) are run and after each permutation a linear regression associated p-value is produced. All the p-values are then plotted and an empirical permuted p-value for each SNP is computed (defined as the probability to obtain a statistic lower than or equal to the one obtained in the first "no permutation" test). The *--qfam-total* procedure is based on the between/within model reported by Fulker et al. (1999) and Abecasis et al. (2000): each genotype score is decomposed in a within family and between-family component, which undergo permutations in the same family *(--qfam-within*) and between different families *(--qfam-between*), respectively. Then the two components are summed to create a new total genotype score, which is tested for association (further details can be found in the PLINK tutorial). An adaptive permutation procedure (*--aperm*) was used, with the following parameters:

- Minimum number of permutations per SNP *1,000*
- Maximum number of permutations per SNP *1,000,000,000*
- Alpha (determining the threshold for pruning p-values) *0*
- Beta (determining the width of confidence interval on empirical p-value) *0.01*
- Initial interval (nr of permutations) to prune SNP test list *100*
- Rate of increase of the initial interval to prune SNP test list *0.001*

Detailed explanations of these arguments can be found in the PLINK tutorial (http://pngu.mgh.harvard.edu/~purcell/plink/).

SNP-based Meta-analysis (METAL)

A sample size-based meta-analysis was run in METAL (Willer et al., 2010). This method consists of computing an overall z-score for each SNP as a weighted sum of z-scores,

determined by the p-value, the direction of the effect and the sample size of each study involved in the meta-analysis.

To this purpose, the default *SCHEME SAMPLESIZE* command was used, along with the *ANALYZE HETEROGENEITY* option to check for the homogeneity of effect sizes across the different datasets (see http://www.sph.umich.edu/csg/abecasis/Metal/index.html for further details).

Gene-based Meta-analysis (VEGAS)

VEGAS performs gene-based association tests for all the 17,787 autosomal genes present in the UCSC Genome Browser map (hg18 assembly), assigning SNPs to genes and combining their effects taking into account the linkage disequilibrium (LD) structure of the genes (Liu et al. 2010). A Bonferroni-corrected significance threshold was set at $p < 2.8 \times 10^{-6}$ to account for the number of genes tested. Gene boundaries were extended up to 50 kb upstream/downstream of 5'-/3'-UTRs, respectively (to include possible variants located in regulatory regions), while the LD patterns for each gene were inferred from the SNP data of the HapMap CEU population (release R2; The International HapMap 3 Consortium, 2010; http://www.hapmap.org/). These options were implemented in the following (default) commands:

-lower 50000 (5'-UTR extension in bp)

-upper 50000 (3'-UTR extension in bp)

-pop HapMap CEU (population of reference for LD inference)

A detailed explanation of the commands is available at http://gump.qimr.edu.au/VEGAS/.

Pathway-based Meta-analysis (INRICH)

The INRICH tool (Lee et al., 2012) for pathway-based association tests takes a set of independent associated genomic intervals and tests them for the enrichment of predefined gene sets (i.e. pathways) through a permutation-based approach. This required extrapolating the associated genomic intervals from the meta-analysis results file through the PLINK *--clump* command, using the following arguments:

--clump-p1 0.001 (p-value threshold for index SNPs = 0.001)

--clump-p2 0.01 (p-value threshold for clumped SNPs = 0.01)

--clump-r2 0.5 (LD (r-squared) threshold for clumping = 0.5)

--clump-kb 250 (physical (kb) threshold for clumping = 250)

Detailed explanations of these arguments can be found in PLINK tutorial (http://pngu.mgh.harvard.edu/~purcell/plink/). Three composite candidate gene sets - representing pathways involved in axon guidance, neuronal migration and steroid sex hormone biology- were tested, using the following INRICH options:

-w 50000 (gene boundaries extension, bp)

-i 10 (minimum nr of genes in tested pathways)

-j 400 (maximum nr of genes in tested pathways)

-p 1 (list gene sets with empirical p-value ≤ 1)

-z 3 (consider only gene sets with ≥ 3 overlapping intervals)

The *INTERVALS* test (default analysis examining enriched association signals for pre-defined sets of genetic variants) and Entrez hg18 gene map for the reference gene file (with genomic coordinates updated to hg19 through the UCSC LiftOver Tool, http://genome.ucsc.edu/cgi-bin/hgLiftOver) were used. Also in this case we extended gene boundaries by 50 kb from the 5'- and 3'-UTRs, to include regulatory regions in the analysis (please notice that the extension is indicated in bp since a bug was found in the current release of the software; Lee, personal communication). For all the other options and commands see the INRICH user manual at http://atgu.mgh.harvard.edu/inrich/started.html.

PLINK Multivariate analysis of top hits

PLINK Multivariate (Ferreira & Purcell, 2009) is a PLINK v1.06 plugin which runs multivariate association tests with several continuous phenotypic traits. Considering a SNP and a set of continuous traits, this tool executes a Canonical Correlation Analysis (CCA), extracting the linear combination of traits that explains the maximum amount of covariation between the SNP and the traits analyzed. This produces as output a single p-value per SNP, representing the significance of multivariate association, and a set of loadings for each trait analyzed, corresponding to the correlation between the trait and the latent variable extracted from all the traits analyzed, and representing the contribution of the trait to the multivariate

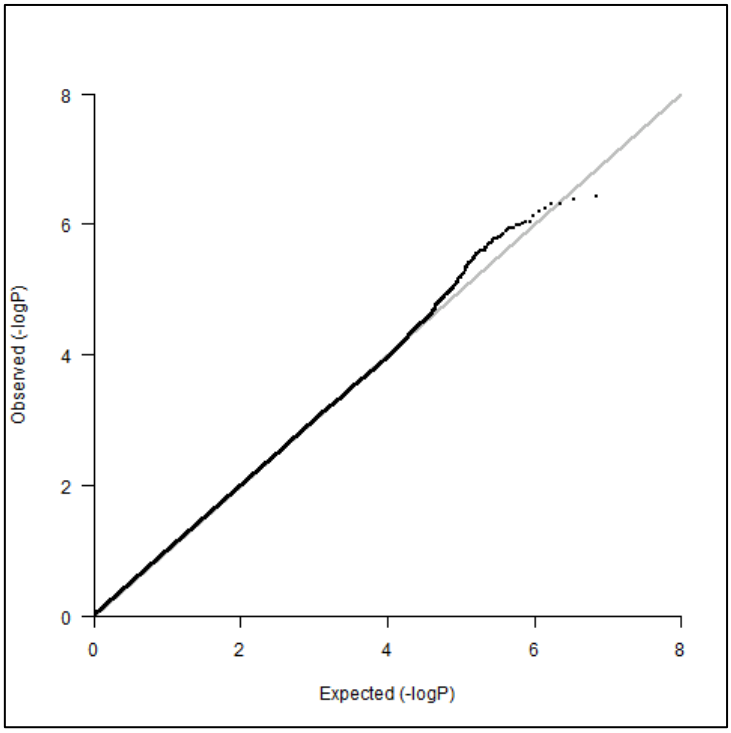
association. See https://genepi.qimr.edu.au/staff/manuelF/multivariate/main.html for detailed explanation.

To adjust for sample relatedness in the datasets, a permutation-based *--mqfam-total* analysis was run. This is a multivariate version of the *–qfam-total* test described above. An adaptive permutation procedure (*--aperm*) was used also in this case, with the same parameters settings used in the univariate QFAM total association test on PC1 scores (see *PLINK QFAM* section above).

S2: Supplementary Results, PC1 meta-analysis

QQ plots

a)



b)

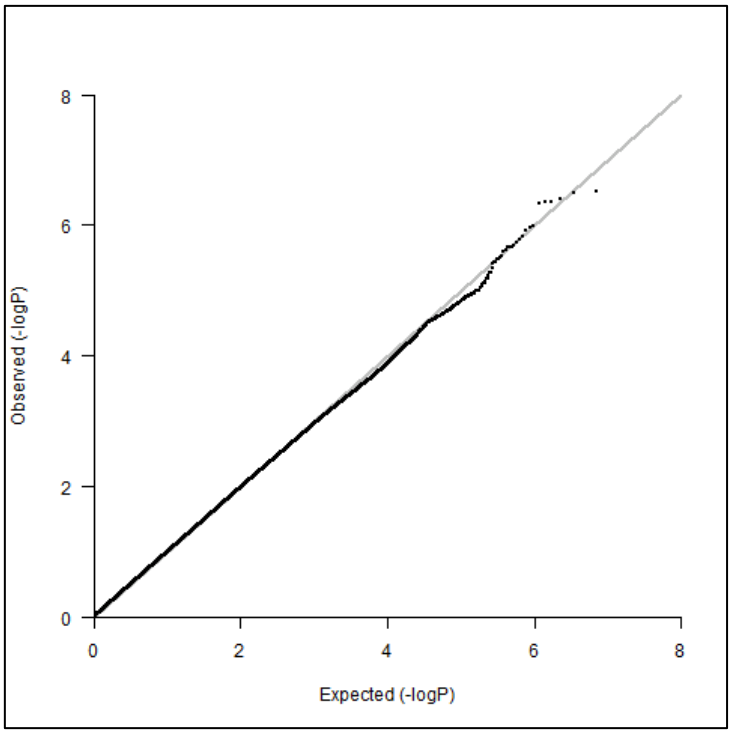
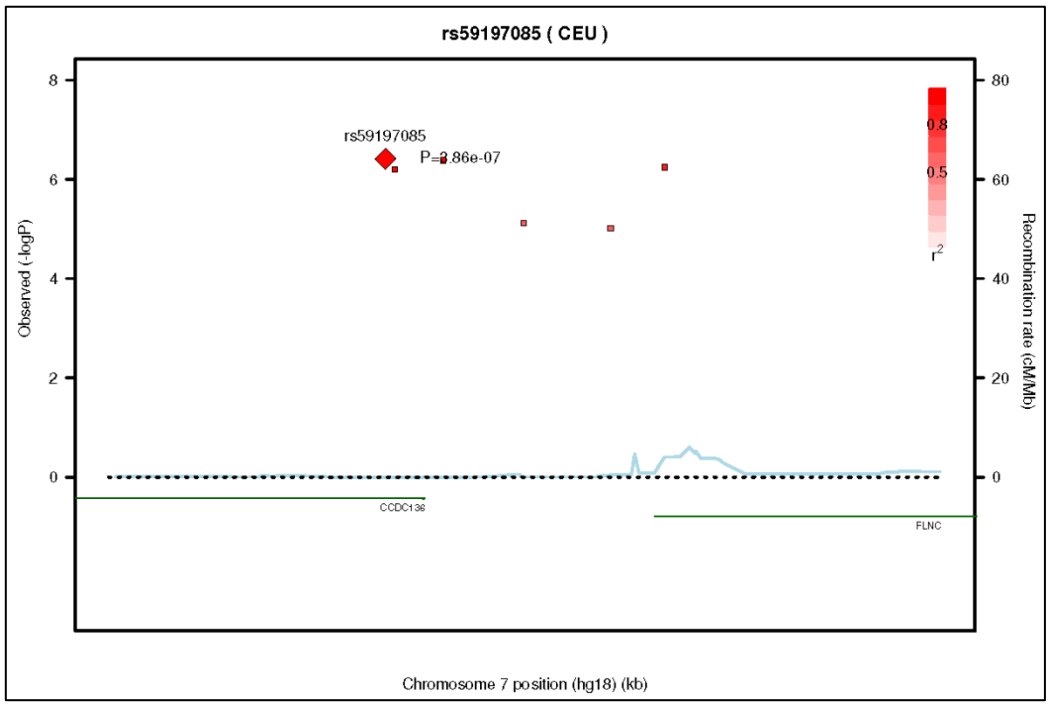


Figure S2 a, b. QQ-plots of the **a)** PC1 and **b)** IQ-adjusted PC1 meta-analyses. The plots were drawn through a dedicated R script (R Core Team, 2013, http://www.r-project.org/).

Association plots

c)



d)

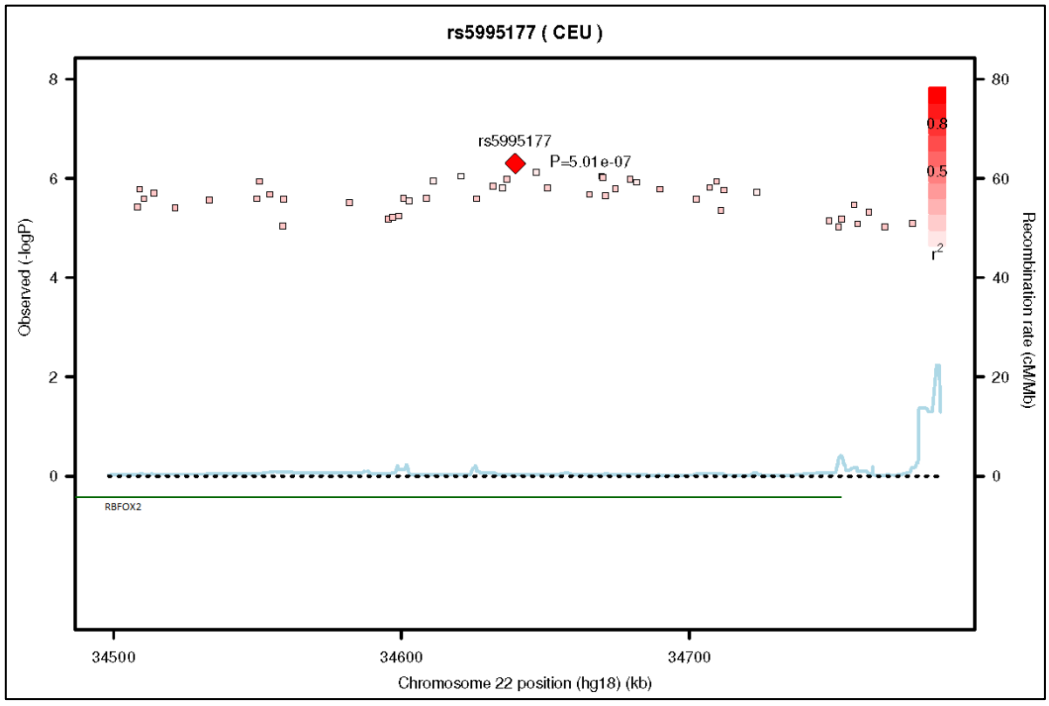


Figure S2 c, d. Association plots of the suggestive hits on **c)** 7q32.1 and **d)** 22q12.3 in the PC1 meta-analysis. All the suggestively associated SNPs ($p < 1x10^{-5}$) are shown. Each squared dot represents an associated SNP in the region and the intensity of color fill represents the level of LD (r^2) with the local top hit (light red indicates low LD, dark red indicates high LD). Note: the plots were produced through the SNAP tool (Johnson et al., 2008; http://www.broadinstitute.org/mpg/snap/ldsearch.php). However, Figure d was slightly modified in order to represent the isoform 5 of *RBFOX2* (the same used for SNP annotation).

S2a)

Chr	SNP[a]	Position (hg19)	Allele1	Allele2	Freq Allele1	Zscore	P-value	Direction[b]	HetPVal[c]	Gene (distance)[d]
7	rs59197085	128460756	a	g	0.08	-5.076	3.86×10^{-7}	----	0.18	OPN1SW(-44.91)|FLNC(-9.726)|CCDC136(0)|CALU(+49.23)|ATP6V1F(-42.14)
7	rs58845495	128462847	t	c	0.92	5.065	4.09×10^{-7}	++++	0.18	OPN1SW(-47)|FLNC(-7.635)|CCDC136(+0.664)|ATP6V1F(-40.05)
7	7:128439695:I	128439695	i	r	0.08	-5.027	4.99×10^{-7}	----	0.21	OPN1SW(-23.85)|FLNC(-30.79)|CCDC136(0)|CALU(+28.17)
22	rs5995177	36309553	a	g	0.08	-5.026	5.01×10^{-7}	----	0.99	RBFOX2(0)
7	rs3734972	128470838	t	c	0.08	-5.003	5.66×10^{-7}	----	0.2	FLNC(0)|CCDC136(+8.655)|ATP6V1F(-32.06)
7	rs3800560	128461094	t	c	0.08	-4.983	6.25×10^{-7}	----	0.17	OPN1SW(-45.25)|FLNC(-9.388)|CCDC136(0)|CALU(+49.57)|ATP6V1F(-41.8)
22	rs12158565	36316843	c	g	0.87	4.946	7.57×10^{-7}	++++	0.44	RBFOX2(0)
22	rs5755979	36290707	t	c	0.13	-4.911	9.05×10^{-7}	----	0.42	RBFOX2(0)
22	rs5750202	36339542	t	c	0.13	-4.911	9.06×10^{-7}	----	0.44	RBFOX2(0)
22	rs5750203	36339998	a	t	0.87	4.897	9.72×10^{-7}	++++	0.4	RBFOX2(0)
22	rs5755990	36306594	t	g	0.13	-4.885	1.03×10^{-6}	----	0.43	RBFOX2(0)
22	rs4541331	36349460	t	c	0.13	-4.884	1.04×10^{-6}	----	0.36	RBFOX2(0)
22	rs5755975	36280999	t	c	0.87	4.869	1.12×10^{-6}	++++	0.48	RBFOX2(0)
22	rs12160116	36379475	t	c	0.13	-4.865	1.15×10^{-6}	----	0.42	RBFOX2(0)
22	rs9619573	36220764	t	c	0.87	4.864	1.15×10^{-6}	++++	0.51	RBFOX2(0)
22	rs5995180	36351751	t	g	0.87	4.856	1.2×10^{-6}	++++	0.47	RBFOX2(0)
22	rs11444345	36393405	i	r	0.13	-4.84	1.3×10^{-6}	----	0.38	RBFOX2(0)
22	rs5755983	36301888	t	c	0.87	4.819	1.44×10^{-6}	++++	0.39	RBFOX2(0)
22	rs5756028	36377043	a	g	0.87	4.809	1.52×10^{-6}	++++	0.34	RBFOX2(0)
22	rs5755987	36305179	t	c	0.87	4.806	1.54×10^{-6}	++++	0.38	RBFOX2(0)
22	rs6000036	36320740	a	g	0.87	4.805	1.54×10^{-6}	++++	0.42	RBFOX2(0)
22	rs5750204	36344302	t	c	0.13	-4.794	1.63×10^{-6}	----	0.48	RBFOX2(0)
22	rs5755948	36179095	a	g	0.13	-4.793	1.64×10^{-6}	----	0.57	RBFOX2(0)
22	rs5756017	36359853	a	g	0.13	-4.791	1.66×10^{-6}	----	0.3	RBFOX2(0)
22	rs5756032	36382102	t	c	0.13	-4.785	1.71×10^{-6}	----	0.27	RBFOX2(0)
22	rs68083039	36248109	i	r	0.13	-4.765	1.89×10^{-6}	----	0.54	RBFOX2(0)
22	rs16996261	36393410	t	g	0.87	4.763	1.91×10^{-6}	++++	0.34	RBFOX2(0)
22	rs5755951	36184094	c	g	0.87	4.757	1.97×10^{-6}	++++	0.57	RBFOX2(0)
22	rs5756005	36335309	a	g	0.13	-4.744	2.09×10^{-6}	----	0.35	RBFOX2(0)
22	rs2092786	36224331	a	t	0.87	4.744	2.1×10^{-6}	++++	0.59	RBFOX2(0)
22	rs5756007	36340848	c	g	0.87	4.731	2.24×10^{-6}	++++	0.4	RBFOX2(0)
22	rs59761754	36204329	i	r	0.13	-4.724	2.32×10^{-6}	----	0.54	RBFOX2(0)
22	rs6000023	36270721	t	c	0.13	-4.708	2.5×10^{-6}	----	0.44	RBFOX2(0)
22	rs5755974	36278740	t	g	0.87	4.707	2.51×10^{-6}	++++	0.35	RBFOX2(0)
22	rs6000004	36180535	a	g	0.87	4.704	2.55×10^{-6}	++++	0.5	RBFOX2(0)
22	rs5995169	36219694	t	g	0.87	4.703	2.56×10^{-6}	++++	0.59	RBFOX2(0)
22	rs5755980	36296128	t	g	0.13	-4.702	2.58×10^{-6}	----	0.44	RBFOX2(0)
22	rs5756023	36372387	a	c	0.13	-4.701	2.59×10^{-6}	----	0.36	RBFOX2(0)
22	rs5750185	36229069	a	g	0.87	4.696	2.65×10^{-6}	++++	0.54	RBFOX2(0)
22	rs5750177	36203266	t	c	0.87	4.692	2.7×10^{-6}	++++	0.52	RBFOX2(0)
22	rs10483192	36272637	a	g	0.13	-4.682	2.84×10^{-6}	----	0.32	RBFOX2(0)
2	rs6737417	222213043	a	g	0.54	4.672	2.98×10^{-6}	++++	0.54	no gene
22	rs113928902	36251888	t	c	0.87	4.663	3.11×10^{-6}	++++	0.5	RBFOX2(0)

Chr	SNP[a]	Position (hg19)	Allele1	Allele2	Freq Allele1	Zscore	P-value	Direction[b]	HetPVal[c]	Gene (distance)[d]
22	rs7289456	36270541	a	g	0.13	-4.65	3.32×10^{-6}	----	0.4	RBFOX2(0)
16	rs28655387	72259192	t	g	0.95	-4.649	3.34×10^{-6}	----	0.42	no gene
22	rs5756049	36427251	t	g	0.13	-4.645	3.39×10^{-6}	----	0.19	RBFOX2(-2.666)
2	rs1025370	222211012	a	g	0.46	-4.638	3.52×10^{-6}	----	0.59	no gene
12	rs10774547	120862716	t	c	0.67	4.632	3.63×10^{-6}	++0+	0.3	TRIAP1(+19.05)|SFRS9(+36.75)|GATC(-21.57)|DYNLL1(-44.94)|COX6A1(-13.19)
22	rs6000002	36178273	t	g	0.87	4.626	3.72×10^{-6}	++++	0.55	RBFOX2(0)
2	rs4674585	222212153	a	c	0.46	-4.62	3.84×10^{-6}	----	0.51	no gene
2	rs11687096	222213174	a	g	0.46	-4.618	3.87×10^{-6}	----	0.62	no gene
22	rs6000006	36191428	a	g	0.87	4.614	3.95×10^{-6}	++++	0.54	RBFOX2(0)
2	rs1025368	222211186	a	g	0.54	4.604	4.15×10^{-6}	++++	0.56	no gene
2	rs6436253	222210670	a	g	0.54	4.598	4.26×10^{-6}	++++	0.59	no gene
22	rs5756031	36380994	t	c	0.13	-4.588	4.47×10^{-6}	----	0.25	RBFOX2(0)
22	rs739200	36432337	a	t	0.89	4.571	4.85×10^{-6}	++++	0.11	RBFOX2(-7.752)
2	rs1036024	222211702	a	c	0.54	4.562	5.07×10^{-6}	++++	0.54	no gene
22	rs9622297	36268975	a	g	0.87	4.537	5.7×10^{-6}	++++	0.47	RBFOX2(0)
20	rs72626581	50783449	t	c	0.75	-4.531	5.88×10^{-6}	----	0.72	ZFP64(0)
2	rs11683727	222210730	c	g	0.46	-4.529	5.94×10^{-6}	----	0.54	no gene
22	rs5750189	36267013	a	c	0.87	4.526	6.02×10^{-6}	++++	0.46	RBFOX2(0)
13	rs141994868	99222422	t	c	0.03	4.508	6.54×10^{-6}	++++	0.81	STK24(0)
22	rs916333	36422904	t	c	0.88	4.507	6.57×10^{-6}	++++	0.18	RBFOX2(0)
22	rs6000021	36265404	t	c	0.87	4.506	6.62×10^{-6}	++++	0.44	RBFOX2(0)
12	rs4766962	120863235	a	t	0.66	4.505	6.62×10^{-6}	++++	0.47	TRIAP1(+18.53)|SFRS9(+36.23)|GATC(-21.05)|DYNLL1(-44.42)|COX6A1(-12.67)
2	rs10498108	222209541	t	c	0.46	-4.503	6.71×10^{-6}	----	0.56	no gene
2	rs1430209	222210212	t	c	0.54	4.496	6.92×10^{-6}	++++	0.54	no gene
22	rs6000066	36418475	t	c	0.12	-4.49	7.12×10^{-6}	----	0.21	RBFOX2(0)
20	rs2038430	50782945	a	g	0.25	4.487	7.23×10^{-6}	++++	0.73	ZFP64(0)
7	rs60894155	128465755	a	g	0.11	-4.475	7.65×10^{-6}	----	0.42	OPN1SW(-49.91)|FLNC(-4.727)|CCDC136(+3.572)|ATP6V1F(-37.14)
22	rs5750221	36447564	a	g	0.87	4.464	8.03×10^{-6}	++++	0.11	RBFOX2(-22.98)
20	20:50780325:I	50780325	i	r	0.25	4.462	8.12×10^{-6}	++++	0.6	ZFP64(0)
22	rs6000071	36428388	t	c	0.13	-4.459	8.24×10^{-6}	----	0.17	RBFOX2(-3.803)
12	rs7970534	120862195	c	g	0.34	-4.452	8.51×10^{-6}	----	0.49	TRIAP1(+19.57)|SFRS9(+37.27)|GATC(-22.09)|DYNLL1(-45.46)|COX6A1(-13.71)
2	rs13384469	7649521	t	g	0.12	4.45	8.59×10^{-6}	++++	0.89	no gene
20	rs6021772	50782343	a	c	0.25	4.447	8.7×10^{-6}	++++	0.65	ZFP64(0)
20	rs4811304	50779338	a	t	0.25	4.438	9.07×10^{-6}	++++	0.69	ZFP64(0)
22	rs5755958	36228855	t	c	0.87	4.438	9.09×10^{-6}	++++	0.69	RBFOX2(0)
22	rs5756045	36421809	c	g	0.88	4.431	9.37×10^{-6}	++++	0.19	RBFOX2(0)
22	rs11089776	36437906	t	g	0.87	4.431	9.4×10^{-6}	++++	0.12	RBFOX2(-13.32)
7	rs3823480	128468881	a	g	0.11	-4.423	9.73×10^{-6}	----	0.44	FLNC(-1.601)|CCDC136(+6.698)|ATP6V1F(-34.02)
2	rs10519830	7657807	a	g	0.88	-4.421	9.81×10^{-6}	----	0.91	no gene
12	rs4767891	120863422	a	g	0.34	-4.421	9.84×10^{-6}	----	0.49	TRIAP1(+18.34)|SFRS9(+36.05)|GATC(-20.86)|DYNLL1(-44.24)|COX6A1(-12.48)
2	rs13429500	7656136	a	t	0.88	-4.419	9.92×10^{-6}	----	0.91	no gene
20	rs58878184	50783204	a	g	0.25	4.418	9.96×10^{-6}	++++	0.69	ZFP64(0)

S2b)

Chr	SNP[a]	Position (hg19)	Allele1	Allele2	Freq Allele1	Zscore	P-value	Direction[b]	HetPVal[c]	Gene (distance)[d]
7	rs59197085	128460756	a	g	0.08	-5.123	3.01×10^{-7}	--+-	0.06	OPN1SW(-44.91)|FLNC(-9.726)|CCDC136(0)|CALU(+49.23)|ATP6V1F(-42.14)
7	rs58845495	128462847	t	c	0.92	5.109	3.23×10^{-7}	++-+	0.06	OPN1SW(-47)|FLNC(-7.635)|CCDC136(+0.664)|ATP6V1F(-40.05)
7	rs3800560	128461094	t	c	0.08	-5.071	3.95×10^{-7}	--+-	0.06	OPN1SW(-45.25)|FLNC(-9.388)|CCDC136(0)|CALU(+49.57)|ATP6V1F(-41.8)
7	7:128439695:I	128439695	i	r	0.08	-5.047	4.48×10^{-7}	--+-	0.07	OPN1SW(-23.85)|FLNC(-30.79)|CCDC136(0)|CALU(+28.17)
7	rs3734972	128470838	t	c	0.08	-5.039	4.68×10^{-7}	--+-	0.06	FLNC(0)|CCDC136(+8.655)|ATP6V1F(-32.06)
9	rs1711745	115452909	a	c	0.09	4.888	1.02×10^{-6}	++++	0.72	KIAA1958(+30.2)|INIP(0)
12	rs10774547	120862716	t	c	0.67	4.852	1.22×10^{-6}	++++	0.48	TRIAP1(+19.05)|SFRS9(+36.75)|GATC(-21.57)|DYNLL1(-44.94)|COX6A1(-13.19)
12	rs4766962	120863235	a	t	0.66	4.79	1.67×10^{-6}	++++	0.59	TRIAP1(+18.53)|SFRS9(+36.23)|GATC(-21.05)|DYNLL1(-44.42)|COX6A1(-12.67)
12	rs4767891	120863422	a	g	0.34	-4.77	1.84×10^{-6}	----	0.66	TRIAP1(+18.34)|SFRS9(+36.05)|GATC(-20.86)|DYNLL1(-44.24)|COX6A1(-12.48)
9	rs1711739	115464474	a	g	0.09	4.745	2.09×10^{-6}	++++	0.87	SNX30(-48.66)|KIAA1958(+41.77)|INIP(0)
9	rs786979	115460453	a	g	0.09	4.737	2.17×10^{-6}	++++	0.83	KIAA1958(+37.75)|INIP(0)
9	rs786978	115460239	a	g	0.1	4.669	3.03×10^{-6}	++++	0.87	KIAA1958(+37.53)|INIP(0)
9	rs786981	115457856	t	g	0.91	-4.666	3.08×10^{-6}	----	0.92	KIAA1958(+35.15)|INIP(0)
9	rs2995805	115454618	t	c	0.91	-4.648	3.35×10^{-6}	----	0.88	KIAA1958(+31.91)|INIP(0)
12	rs7970534	120862195	c	g	0.34	-4.633	3.6×10^{-6}	----	0.6	TRIAP1(+19.57)|SFRS9(+37.27)|GATC(-22.09)|DYNLL1(-45.46)|COX6A1(-13.71)
9	rs1418410	115465282	a	g	0.91	-4.627	3.71×10^{-6}	----	0.78	SNX30(-47.85)|KIAA1958(+42.58)|INIP(0)
9	rs786983	115458629	t	c	0.91	-4.622	3.8×10^{-6}	----	0.88	KIAA1958(+35.92)|INIP(0)
9	rs2798316	115462552	t	c	0.9	-4.587	4.49×10^{-6}	----	0.85	KIAA1958(+39.85)|INIP(0)
9	rs2185768	115462645	t	c	0.1	4.556	5.21×10^{-6}	++++	0.87	KIAA1958(+39.94)|INIP(0)
16	rs28655387	72259192	t	g	0.95	-4.478	7.53×10^{-6}	----	0.26	no gene
9	rs1711744	115453484	a	t	0.09	4.478	7.53×10^{-6}	++++	0.85	KIAA1958(+30.78)|INIP(0)
9	rs1965335	115504483	t	c	0.09	4.472	7.74×10^{-6}	++++	0.63	SNX30(-8.65)|INIP(+24.1)
9	rs72768411	115502670	a	c	0.09	4.46	8.2×10^{-6}	++++	0.68	SNX30(-10.46)|INIP(+22.28)
2	rs6737417	222213043	a	g	0.54	4.452	8.51×10^{-6}	++++	0.54	no gene
12	rs11065109	120863914	t	c	0.66	4.42	9.87×10^{-6}	++++	0.63	TRIAP1(+17.85)|SFRS9(+35.56)|GATC(-20.37)|DYNLL1(-43.74)|COX6A1(-11.99)
7	rs55907818	128477620	a	g	0.08	-4.418	9.98×10^{-6}	--++	0.09	FLNC(0)|CCDC136(+15.44)|ATP6V1F(-25.28)

Table S2. Top associations (p-values < $1x10^{-5}$) of the SNP-based **a)** PC1 and **b)** IQ-adjusted PC1 GWAS meta-analysis. [a] Single-base indels were not filtered out from imputed markers since they were reliably called in the imputation reference (1000 Genomes, Phase I v3), and were tested for association as they could represent coding frameshift variants of biological interest. [b] The direction of effect of Allele1 is reported for datasets in the following order: CLDRC-RD, UK-RD, SLIC, CLDRC-ADHD. [c] Test for the homogeneity of effect sizes across the different datasets ($p \geq 0.05$ indicates homogeneous effects). [d] Physical distance (kb) from close genes (in a ±50kb range from each marker) is indicated, along with orientation based on the direction of transcription ("-" = upstream of 5'-UTR, "+" = downstream of 3'-UTR).

S2c)

Chr	Gene	nSNPs	Start[a]	Stop[a]	Pvalue	Best-SNP[b]	SNP-pvalue
22	RBFOX2	141	34464728	34754531	5×10^{-6}	rs12158565	7.57×10^{-7}
5	ZNF346	53	176382302	176426351	2.8×10^{-5}	rs6874617	5.92×10^{-5}
7	FLNC	52	128257718	128286564	4.6×10^{-5}	rs3734972	5.66×10^{-7}
5	FGFR4	34	176446526	176457732	4.9×10^{-5}	rs7707602	8.86×10^{-5}
7	ATP6V1F	40	128290133	128293138	1.19×10^{-4}	rs3734972	5.66×10^{-7}
9	INIP	89	114488611	114520208	1.5×10^{-4}	rs1711745	1.31×10^{-5}
5	UIMC1	108	176264611	176366049	2.66×10^{-4}	rs6874617	5.92×10^{-5}
1	SLC16A4	109	110707027	110735159	2.93×10^{-4}	rs11102092	2.65×10^{-5}
11	LRFN4	30	66381451	66384522	3.71×10^{-4}	rs7948839	1.85×10^{-4}
7	CCDC136	69	128219334	128249419	3.79×10^{-4}	rs3734972	5.66×10^{-7}
6	MOXD1	162	132658886	132764357	4.06×10^{-4}	rs7450274	2.01×10^{-4}
20	ZFP64	183	50133956	50241931	4.88×10^{-4}	rs2038430	7.23×10^{-6}
9	SNX30	254	114552954	114677088	4.97×10^{-4}	rs1418410	3.02×10^{-5}
11	PC	49	66372572	66482423	6.23×10^{-4}	rs7948839	1.85×10^{-4}
1	RBM15	108	110683467	110690826	6.35×10^{-4}	rs11102092	2.65×10^{-5}
11	RCE1	24	66367458	66370579	6.54×10^{-4}	rs7948839	1.85×10^{-4}
12	COX6A1	44	119360286	119362912	6.64×10^{-4}	rs4766962	6.62×10^{-6}
9	SLC46A2	184	114681020	114692866	7.05×10^{-4}	rs1475293	8.46×10^{-5}
17	SP2	88	43328514	43361322	7.76×10^{-4}	rs3096	6.93×10^{-4}
6	THEM2	136	24775253	24809921	7.96×10^{-4}	rs7768291	1.13×10^{-4}
12	GATC	57	119368666	119382145	7.98×10^{-4}	rs4766962	6.62×10^{-6}
4	BTC	92	75890471	75938906	8.15×10^{-4}	rs7667066	4.91×10^{-5}
12	SFRS9	58	119383853	119391941	8.66×10^{-4}	rs4766962	6.62×10^{-6}
7	AKR1B10	70	133862938	133876700	8.75×10^{-4}	rs1732049	2.12×10^{-4}
12	TRIAP1	48	119366146	119368598	9×10^{-4}	rs4766962	6.62×10^{-6}
17	TMEM105	41	76899668	76919069	9.01×10^{-4}	rs7219316	3.29×10^{-4}
17	C17orf55	48	76891218	76897643	9.29×10^{-4}	rs7219316	3.29×10^{-4}
1	HBXIP	102	110745399	110752069	9.3×10^{-4}	rs11102092	2.65×10^{-5}
8	KIAA1833	14	145274906	145388831	9.7×10^{-4}	rs11989162	1.01×10^{-3}

S2d)

Chr	Gene	nSNPs	Start[a]	Stop[a]	Pvalue	Best-SNP[b]	SNP-pvalue
9	INIP	89	114488611	114520208	3.4×10^{-5}	rs1711745	1.02×10^{-6}
12	GATC	57	119368666	119382145	6.6×10^{-5}	rs4766962	1.67×10^{-6}
12	COX6A1	44	119360286	119362912	6.9×10^{-5}	rs4766962	1.67×10^{-6}
12	TRIAP1	48	119366146	119368598	8.9×10^{-5}	rs4766962	1.67×10^{-6}
12	SFRS9	58	119383853	119391941	9.3×10^{-5}	rs4766962	1.67×10^{-6}
9	SNX30	254	114552954	114677088	1.11×10^{-4}	rs1711739	2.09×10^{-6}
7	FLNC	52	128257718	128286564	1.21×10^{-4}	rs3734972	4.68×10^{-7}
5	FGFR4	34	176446526	176457732	1.26×10^{-4}	rs6861120	1.98×10^{-4}
5	ZNF346	53	176382302	176426351	1.28×10^{-4}	rs11954635	1.22×10^{-4}
9	SLC46A2	184	114681020	114692866	1.28×10^{-4}	rs1324930	1.38×10^{-5}
12	DYNLL1	72	119392042	119420681	1.4×10^{-4}	rs4766962	1.67×10^{-6}
7	ATP6V1F	40	128290133	128293138	2.14×10^{-4}	rs3734972	4.68×10^{-7}
22	RBFOX2	141	34464728	34754531	2.19×10^{-4}	rs12160116	4.04×10^{-5}
12	COQ5	71	119425464	119451347	2.6×10^{-4}	rs3213565	5.9×10^{-4}
12	RNF10	60	119456514	119499780	3.26×10^{-4}	rs3213565	5.9×10^{-4}
4	BTC	92	75890471	75938906	3.87×10^{-4}	rs7667066	4.02×10^{-5}
12	POP5	39	119501230	119503584	5.16×10^{-4}	rs3213565	5.9×10^{-4}
5	UIMC1	108	176264611	176366049	6.34×10^{-4}	rs11954635	1.22×10^{-4}
11	LRFN4	30	66381451	66384522	7.35×10^{-4}	rs2167457	2.85×10^{-4}
7	CCDC136	69	128219334	128249419	7.41×10^{-4}	rs3734972	4.68×10^{-7}
17	SP2	88	43328514	43361322	7.45×10^{-4}	rs11079803	1.01×10^{-3}
6	MOXD1	162	132658886	132764357	8.24×10^{-4}	rs17792959	3.28×10^{-4}
1	IVNS1ABP	46	183532144	183553084	8.67×10^{-4}	rs6689206	2.76×10^{-4}

Table S2. Top hits of the gene-based (VEGAS) association tests (significance: $p < 2.8 \times 10^{-6}$) in the **c)** PC1 and **d)** IQ-adjusted PC1 GWAS meta-analysis. Only genes with p-values < 0.001 are reported. [a] Start and stop positions are expressed in hg18 coordinates. [b] Note that this analysis was based on SNPs included in HapMap II CEU reference, therefore the most significantly associated SNP does not necessarily coincide with other analyses based on imputation with the 1000 Genomes reference dataset.

S2e)

Candidate pathway	Pathway size (nr of genes)	Overlaps	Empirical P	Corrected P
axonal guidance[a]	89	13	0.03	0.071
neuronal migration[b]	64	7	0.196	0.405
steroids[c]	333	25	0.023	0.051

S2f)

Candidate pathway	Pathway size (nr of genes)	Overlaps	Empirical P	Corrected P
axonal guidance[a]	89	10	0.182	0.423
neuronal migration[b]	64	6	0.314	0.61
steroids[c]	333	23	0.041	0.097

Table S2. Pathway-based (INRICH) analysis of association signals detected in **e)** PC1 and **f)** IQ-adjusted PC1 meta-analysis. Significance: corrected $p < 0.05$. [a] All the GO sets containing the term "axon guidance". [b] All the GO sets containing the term "neuron migration".[c] All the GO sets containing the terms "steroid", "androgen", "estrogen", "progesterone" and "testosterone".

S2g)

Trait	PC1			IQ-adjusted PC1		
Dataset	P-value	Weighted Z score	Beta[a]	P-value	Weighted Z score	Beta[a]
CLDRC-RD	9.11×10^{-3}	-1.41	-0.311	5.9×10^{-3}	-1.5	-0.315
UK-RD	1.21×10^{-6}	-3.4	-0.435	3.67×10^{-7}	-3.53	-0.436
SLIC	0.707	-0.14	-0.072	0.877	0.06	0.027
CLDRC-ADHD	0.655	-0.13	-0.119	0.606	-0.15	-0.13
Meta-Analysis	3.86×10^{-7}	-5.08	NA[b]	3.01×10^{-7}	-5.12	NA[b]

S2h)

Trait	PC1			IQ-adjusted PC1		
Dataset	P-value	Weighted Z score	Beta[a]	P-value	Weighted Z score	Beta[a]
CLDRC-RD	5.61×10^{-3}	-1.5	-0.307	0.021	-1.26	-0.234
UK-RD	2.78×10^{-4}	-2.54	-0.348	1.16×10^{-3}	-2.25	-0.305
SLIC	0.072	-0.66	-0.34	0.232	-0.44	-0.197
CLDRC-ADHD	0.26	-0.33	-0.204	0.195	-0.38	-0.233
Meta-Analysis	5.01×10^{-7}	-5.03	NA[b]	1.5×10^{-5}	-4.33	NA[b]

Table S2. Contribution of each GWAS to the strength of the association in the PC1 and IQ-adjusted PC1 meta-analysis, for the top association signals **g)** rs59197085 (7q32.1) and h**)** rs5995177 (22q12.3). These are represented by PLINK univariate QFAM p-values and beta regression coefficients for each GWAS, and by corresponding weighted Z-scores, as computed by METAL sample size based algorithm (Willer et al. 2010). The sign of z scores and beta values refer to the allelic trend of the minor allele (A in both cases). [a] Although beta values computed by QFAM are not adjusted for family structure, they are reported in the table as a term of comparison for effect sizes. [b] Not Applicable, since the METAL sample size based algorithm computes a global weighted Z score (but not a Beta coefficient).

S3: Supplementary Results, $PC1_{read}$ meta-analysis

Table S3. Top association signals ($p < 1x10^{-5}$) in the **a)** $PC1_{read}$ and **b)** IQ-adjusted $PC1_{read}$ GWAS meta-analysis. [a] The direction of effect of Allele1 is reported for datasets in the following order: CLDRC-RD, UK-RD, SLIC, CLDRC-ADHD. [b] Physical distance (kb) from closest genes (in a ±10kb range from each marker) is indicated, along with orientation based on the direction of transcription ("-" = upstream of 5'-UTR, "+" = downstream of 3'-UTR).

S3a)

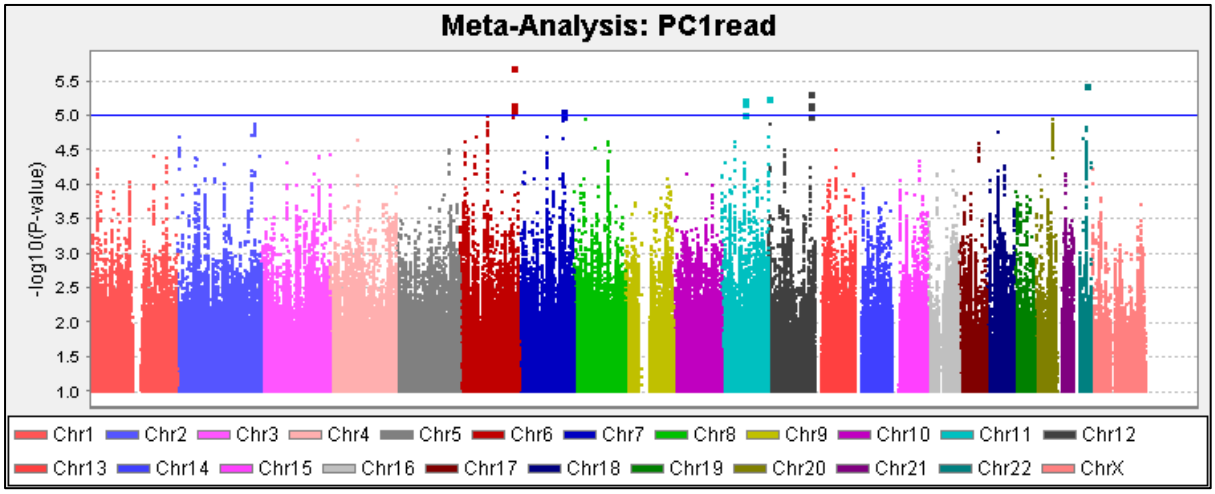
Chr	SNP	Position (hg19)	Allele1	Allele2	Freq Allele1 (%)	P-value	Direction[a]	Gene (distance)[b]	Variant type
6	rs56139919	155914380	a	g	1.806	2.02×10^{-6}	+++-	no gene	
22	rs5995177	36309553	a	g	8.049	3.56×10^{-6}	----	RBFOX2(0)	intronic
12	rs10774547	120862716	t	c	67.02	4.76×10^{-6}	++-+	no gene	
11	rs118151645	133866086	c	g	96.437	5.57×10^{-6}	++++	no gene	
11	rs2275998	66326581	t	c	80.97	5.89×10^{-6}	----	CTSF(+4.353kb)|ACTN3(0)	intronic
11	rs2229455	66328055	a	g	80.97	6.64×10^{-6}	----	CTSF(+2.879kb)|ACTN3(0)	exonic, synonymous
12	rs4766962	120863235	a	t	66.07	6.99×10^{-6}	++-+	no gene	
6	rs7765720	155830568	c	g	98.102	7.03×10^{-6}	---+	no gene	
12	rs7970534	120862195	c	g	33.94	7.33×10^{-6}	--+-	no gene	
6	rs113262260	155854928	a	g	1.899	8.16×10^{-6}	+++-	no gene	
7	rs3800560	128461094	t	c	7.971	8.46×10^{-6}	----	FLNC(-9.388)|CCDC136(0)	intronic
7	rs58845495	128462847	t	c	92.029	9.02×10^{-6}	++++	FLNC(-7.635)|CCDC136(+0.664)	
11	rs2229456	66328741	a	c	80.97	9.32×10^{-6}	----	CTSF(+2.193kb)|ACTN3(0)	exonic, missense
7	rs59197085	128460756	a	g	7.971	9.78×10^{-6}	----	FLNC(-9.726)|CCDC136(0)	intronic
12	rs4767891	120863422	a	g	33.93	9.91×10^{-6}	--+-	no gene	

S3b)

Chr	SNP	Position (hg19)	Allele1	Allele2	Freq Allele1 (%)	P-value	Direction[a]	Gene (distance)[b]	Variant type
11	rs1496243	133620944	a	g	33.34	3.41×10^{-7}	----	no gene	
11	rs4937830	133645903	c	g	64.82	5.53×10^{-7}	++++	no gene	
11	rs10894745	133647016	a	g	34.77	1.21×10^{-6}	----	no gene	
11	rs7944602	133612417	a	g	33.48	1.32×10^{-6}	----	no gene	
11	rs4936208	133644469	t	c	34.77	1.38×10^{-6}	----	no gene	
11	rs4937829	133642497	a	g	34.73	1.56×10^{-6}	----	no gene	
11	rs2220960	133639946	a	g	65.3	1.87×10^{-6}	++++	no gene	
6	rs56139919	155914380	a	g	1.806	2.23×10^{-6}	+++-	no gene	
12	rs10774547	120862716	t	c	67.02	3.39×10^{-6}	++-+	no gene	
11	rs4936207	133631687	a	c	35.66	3.45×10^{-6}	----	no gene	
12	rs4766962	120863235	a	t	66.07	3.60×10^{-6}	++++	no gene	
11	rs10431101	133615843	t	c	65.25	3.65×10^{-6}	++++	no gene	
11	rs6590728	133618314	t	c	34.52	3.79×10^{-6}	----	no gene	
12	rs7970534	120862195	c	g	33.94	4.03×10^{-6}	----	no gene	
12	rs4767891	120863422	a	g	33.93	4.32×10^{-6}	----	no gene	
7	rs58845495	128462847	t	c	92.029	6.54×10^{-6}	++-+	FLNC(-7.635)|CCDC136(+0.664)	
6	rs16890716	80131140	a	g	84.71	6.57×10^{-6}	++++	no gene	
6	rs17800074	80126873	t	c	15.21	7.26×10^{-6}	----	no gene	
6	rs62411317	80128434	a	c	84.76	7.38×10^{-6}	++++	no gene	
7	rs3800560	128461094	t	c	7.971	8.56×10^{-6}	--+-	FLNC(-9.388)|CCDC136(0)	intronic
6	rs62411314	80120593	a	t	15.19	8.71×10^{-6}	----	no gene	
7	rs59197085	128460756	a	g	7.971	9.11×10^{-6}	--+-	FLNC(-9.726)|CCDC136(0)	intronic

Manhattan plots

a)



b)

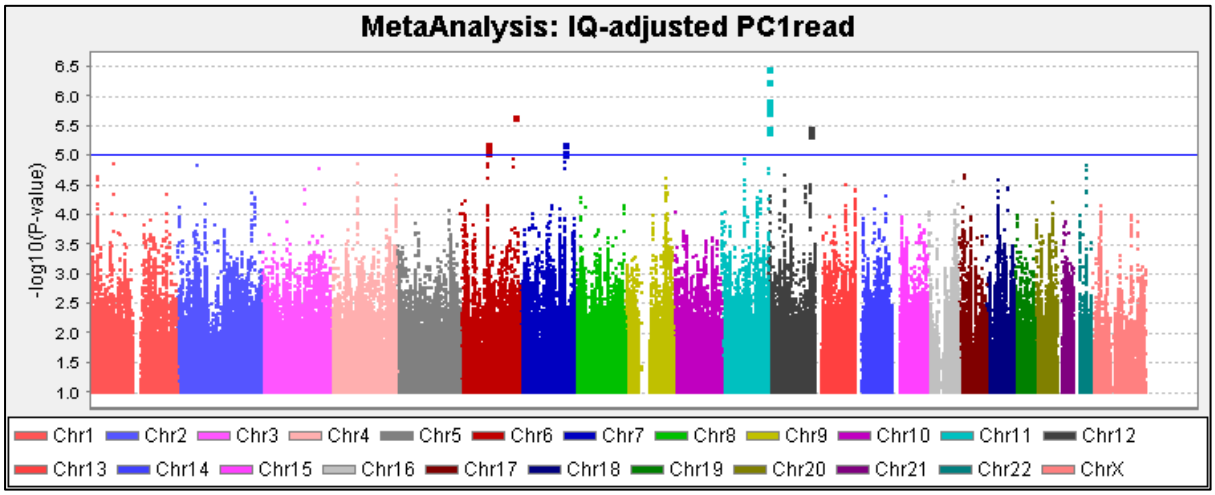
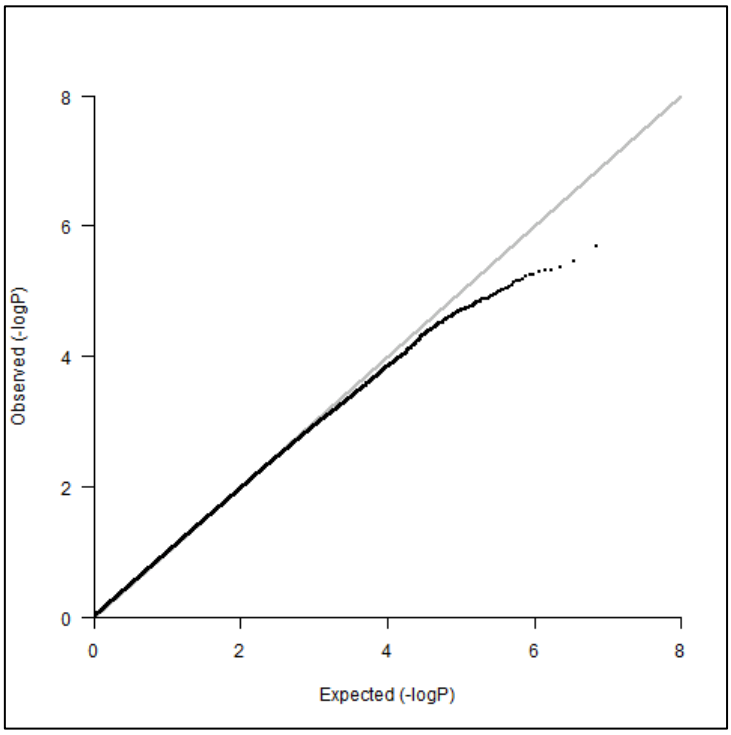


Figure S3. Manhattan plots of the **a)** $PC1_{read}$ and **b)** IQ-adjusted $PC1_{read}$ genome-wide association scan meta-analysis. The blue lines represent the nominal suggestive significance threshold ($p = 1x10^{-5}$).

QQ plots

c)



d)

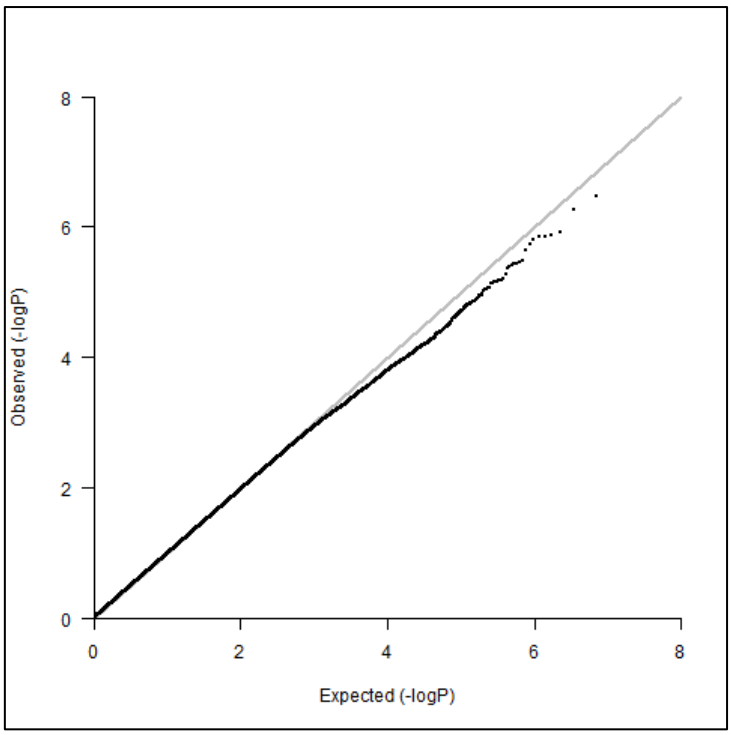


Figure S3. QQ-plots of the **c)** $PC1_{read}$ and **d)** IQ-adjusted $PC1_{read}$ genome-wide association scan meta-analysis. The plots were drawn through a dedicated R script (R core Team 2013, http://www.r-project.org/).

Chapter 4:
Testing associations of candidate SNPs and genes previously implicated in reading and language: the consistency issue

This chapter is based on:

Gialluisi, A., Newbury, D.F., Olson, R.K., DeFries, J.C., Paracchini, S., Francks, C. & Fisher, S.E. Testing associations of candidate SNPs and genes previously implicated in reading and language: the consistency issue. (*in prep*)

Abstract

Reading Disability (RD) and Specific Language Impairment (SLI) are comorbid disorders which are thought to have shared genetic underpinnings. So far, a relatively small number of genes have been implicated in RD and/or SLI in more than one dataset. In the present chapter we assess the association of single-nucleotide polymorphisms (SNPs) in these candidate genes with a principal component (PC) score derived from several reading and language traits, in a meta-analysis of >1,800 subjects.

Our investigation focused on 30 candidate polymorphisms. 25 of these had been investigated in subsets of our datasets in earlier studies. For these SNPs our main goal was to provide an overview of pleiotropic associations, rather than independent replications of previous findings. In this category of SNPs, we observed significant associations with PC scores ($p \sim 10^{-2}$-10^{-4}) for rs2143340, rs3212236, rs9461045 and rs761100 in *KIAA0319* (6p22.3); and for rs16973771, rs2875891 and rs8045507 in *ATP2C2* (16q24.1). For 5 SNPs which had been originally implicated in different datasets, we attempted independent replications of association. Among the latter, rs12495133 in *ROBO1* (3p12) -recently found to be associated with RD- was significantly associated with PC scores ($p \sim 10^{-4}$). All these associations showed directions of effect consistent with those of the original studies, and effects on various reading and language traits, as revealed by univariate and multivariate association testing.

Gene-based analysis of the same candidate genes revealed significant associations of *KIAA0319* and *ROBO1*, in line with the results of the SNP-based assessment.

This study supports the influence of *KIAA0319* and *ROBO1* on both reading- and language-related phenotypes, providing independent statistical support for the association at rs12495133 (*ROBO1*). On the other hand, the lack of replication for other candidate SNPs casts doubt on the replicability of the original findings.

Introduction

Reading disability (RD, or developmental dyslexia) and Specific Language Impairment (SLI) are deficits in acquiring normal reading and language skills, respectively, in spite of the absence of overt reasons such as neurological deficits, low intelligence, or inadequate socioeconomic and educational opportunity (Shaywitz et al., 1990; Bishop, 1994).

RD and SLI are two of the most prevalent neurodevelopmental disorders in school-aged populations, with prevalence estimates of ~5-8% (Shaywitz et al., 1990; Tomblin et al., 1997). They are etiologically complex, with a strong genetic basis (Bishop & Snowling, 2004). Both RD/SLI and continuous reading-/language-related traits show moderate to high heritabilities (30-70%; Bishop & Hayiou-Thomas, 2008; Fisher & DeFries, 2002; see Chapter 2 for an overview of heritability studies). RD and SLI are also frequently comorbid, with comorbidity rates ranging between 43% and 55% (Snowling et al. 2000; McArthur et al., 2000).

So far, various different genes have been tentatively associated with RD, SLI and/or reading/language traits, mainly through linkage studies followed by candidate gene association studies (reviewed in Carrion-Castillo et al., 2013; Raskind et al., 2013; Newbury et al., 2010; Newbury & Monaco, 2010). Only a few of these genes have been repeatedly implicated in reading and language, by at least two independent studies or by analyses of independent datasets in the same study. These include *DYX1C1* (15q21; Taipale et al., 2003), *KIAA0319* (6p22; Francks et al., 2004; Cope et al., 2005; Harold et al., 2006; Dennis et al., 2009), *DCDC2* (6p22; Meng et al., 2005; Schumacher et al., 2006), *MRPL19/GCFC2* (2p12; Anthoni et al., 2007) and *ROBO1* (3p12, Hannula-Jouppi et al., 2005; Bates et al.; 2011; Tran et al., 2014) for RD and reading-related traits; and *FOXP2* (7q31; Fisher & Scharff, 2009), *CNTNAP2* (7q35; Vernes et al., 2008), *CMIP* and *ATP2C2* (16q23 and 16q24; Newbury et al., 2009) for SLI and relevant language traits.

Overview of candidate RD/SLI genes

For the purposes of this chapter, we review below the genes most consistently implicated in RD and/or SLI and the specific genetic variants associated, which mainly include Single Nucleotide Polymorphisms (SNPs), point mutations and structural rearrangements. Several independent studies had previously detected linkage of these regions to RD/SLI. An overview

of these studies can be found in Chapter 1 and, in more detail, elsewhere (Carrion-Castillo et al., 2013; Raskind et al., 2013; Newbury et al., 2010; Newbury & Monaco, 2010).

DYX1C1 (15q21.3)

Dyslexia susceptibility 1 candidate 1 *(DYX1C1)* was the first gene to be identified in RD etiology. This was first detected in a Finnish family where a balanced translocation t(2;15)(q11;q21) co-segregated with reading difficulties (Nopola-Hemmi et al., 2000). The breakpoint on chromosome 15 was located within the *DYX1C1* gene, in a region which had been consistently liked to dyslexia and reading traits (Taipale et al., 2003). Since then, significant associations with RD, reading traits or other related cognitive abilities have been reported for many SNPs at this locus (see Carrion-Castillo et al., 2013 for a review). Most prominently, rs57809907 and rs3743205 were found to be associated with RD in two Finnish family-based datasets (N ~170 and ~140), in a classical case-control association study (Taipale et al., 2003). Nonetheless, other studies were not able to replicate these associations, or reported opposite directions of effect (Carrion-Castillo et al., 2013). Meta-analyses of several association studies on rs57809907 and rs3743205 provided weak or no evidence of association with dyslexia risk (Zou et al., 2012; Tran et al., 2013). *DYX1C1* encodes a product involved in protein-protein interaction (Taipale et al., 2003) and has been suggested to play a role in neuronal migration, Central Nervous System (CNS) development and cell-cell adhesion (Adler et al., 2013; Tammimies et al., 2013). A growing body of molecular research supports the involvement of Dyx1c1 also in cilia function and assembly (Chandrasekar et al., 2013; Tarkar et al., 2013).

KIAA0319 (6p22.3)

KIAA0319 is one of the candidates most consistently implicated in RD. Francks et al. (2004) reported for the first time a specific association at this locus, through analysis of several reading-related traits in two family-based datasets partially overlapping with the CLDRC-RD (N=369) and UK-RD (N=630) datasets analyzed in this study. More specifically, significant association was reported with a ~77 kb haplotype, tagged by the SNPs rs4504469, rs2038137 and rs2143340. This haplotype overlapped the first four exons of *KIAA0319*, spanned *TDP2* (tyrosyl-DNA phosphodiesterase 2, also known as *TTRAP*, TRAF and TNF receptor associated protein) and was located just upstream of *ACOT13* (acyl-CoA thioesterase 13, also

known as *THEM2*, thioesterase superfamily member 2). Another study reported an association with categorical dyslexia for a partially overlapping haplotype, rs4504469-rs6935076, in a case-control study (N ~500) and in a semi-independent sample of 143 RD trios (Cope et al., 2005). A candidate SNP association study on two RD datasets from the UK -overlapping with those used by Francks et al. (2004) and Cope et al. (2005) and testing more SNPs in the 6p22 region- partly failed to replicate these findings (Harold et al., 2006). However, it reported two interesting associations with RD in the putative promoter region of *KIAA0319*, namely at rs3212236 and rs761100. Another significant association with several reading traits was reported for rs9461045, a SNP lying in the original 77 kb haplotype discovered by Francks et al. (2004), which was also associated with reduced expression of the gene in neuronal cells (Dennis et al., 2009). Many of the polymorphisms mentioned above were found to be associated with continuous reading-related traits also in large population-based datasets, both at the single marker and at the haplotype level (Luciano et al., 2007; Paracchini et al., 2008; Scerri et al., 2011). However, direction of effects on these traits were not always concordant with the original studies, while some other studies have not reported any evidence of association at these SNPs (reviewed in Carrion-Castillo et al., 2013). A meta-analysis of several association studies tried to clarify these inconsistencies, reporting a significant association with RD for the minor allele of rs4504469 (Zou et al., 2012). SNPs in *KIAA0319* have also been associated with SLI (Rice et al., 2009), expressive/receptive language and nonword repetition (Newbury et al., 2011), suggesting potential pleiotropic effects of this gene on language-related deficits. Although the functional characterization of KIAA0319 is still far from being clear, this protein was hypothesized to be involved in cell-cell adhesion and interaction during neuronal migration in the developing CNS (Velayos-Baeza et al., 2007; 2008). Consistent with this view, knockdown of Kiaa0319 in embryonic rat neocortex leads to disruptions in neuronal migration, periventricular heterotopia and structural defects in dendrites (Peschansky et al., 2010; Adler et al., 2013). However, caution is needed in the interpretation of these studies, since off-target effects have been reported for short hairpin RNAs (shRNAs), normally used for RNA interference (Baek et al., 2014).

DCDC2 (6p22.3)

In addition to *KIAA0319*, another gene in the 6p22.3 region, *DCDC2* (doublecortin domain containing 2), has been implicated in RD etiology. Meng and colleagues (2005) reported the association of a 2.4 kb deletion at this locus with orthographic coding, in a US family-based

sample (N=536), partially overlapping with the CLDRC-RD dataset used in our study. This deletion was in combination with a compound Short Tandem Repeat (STR) called BV677278, which was later shown to be associated with dyslexic status (Schumacher et al., 2006). More recently, association with a quantitative reading trait has also been reported for the above mentioned deletion (Marino et al., 2012). Molecular analyses of BV677278 detected a binding site for a transcription factor expressed in the human brain, ETV6, and reported an effect of the different STR alleles on *DCDC2* expression in the CNS (Powers et al., 2013; Meng et al., 2011). However, other studies, such as Harold et al. (2006), reported only weak or inconsistent associations for this variant. Similarly, SNPs in high LD with BV677278 recently showed inconsistent associations with both categorical RD/SLI and continuous reading and language traits (Powers et al., 2013; Eicher et al., 2014). Additional SNP markers in *DCDC2* have been associated with dyslexia: rs807724 and rs1087266 have been reported to be associated with a composite score of word reading, spelling and comprehension in the CLDRC sample (Meng et al., 2005); while rs793862, rs807701 and the resulting haplotype were significantly overtransmitted to RD cases in two independent German samples (137 and 239 dyslexic trios; Schumacher et al., 2006). These associations were replicated, both with categorical dyslexia (Wilcke et al., 2009; Newbury et al., 2011) and with continuous reading-related skills such as phoneme awareness (Harold et al. 2006), word reading and nonword repetition (Scerri et al., 2011). As for other candidate genes, other studies failed to replicate these findings (reviewed in Carrion-Castillo et al., 2013) and meta-analyses were able to demonstrate global evidence of association with RD only for one of these candidate SNPs, rs807701 (Zhong et al., 2013). *DCDC2* encodes a microtubule-binding protein which is thought to have a role in primary cilia structure and signaling (Massinen et al., 2011; Grati et al., 2015; Schueler et al., 2015). A prominent role of Dcdc2 in neuronal migration and dendrite outgrowth has also been proposed by gene knockdown studies (Meng et al., 2005; Wang et al., 2011). As above, caution is suggested in the interpretation of these results, not only for the off-target effects sometimes produced by RNA interference (Baek et al., 2014), but also because Dcdc2 knockout models do not show any anomaly in these processes (Wang et al., 2011).

MRPL19/GCFC2 (2p12)

The involvement of *GCFC2* (GC-rich sequence DNA-binding factor 2, also known as *C2ORF3*, chromosome 2 open reading frame 3) and *MRPL19* (mitochondrial ribosomal

protein L19) in RD etiology is supported by a single study which provided evidence of association for the 2p12 region (Anthoni et al., 2007). In this combined linkage/association study, both single SNP and haplotype significant associations were first reported in a set of 11 dyslexic Finnish pedigrees, implicating the SNPs rs917235 and rs730148 and the haplotype rs10000585-rs917235-rs714039. In an independent sample of 251 dyslexic German families, another significant haplotype association was found with RD for the three-markers haplotype rs917235-rs714939-rs6732511, partially overlapping the haplotype detected in the discovery sample and with concordant direction of effect. These haplotype associations were confirmed in a joint analysis of the two sample sets, covering a total of 16.6 kb, in a region not far from the genes *MRPL19* and *GCFC2*. Heterozygous carriers of the associated risk haplotypes showed attenuated expression of both *MRPL19* and *GCFC2* compared with non-carriers (Anthoni et al., 2007). Nonetheless, the putative role of these genes in RD etiology is mostly unknown and other candidate SNP studies did not provide any replication of these associations, neither with dyslexic status (Venkatesh et al., 2013) nor with continuous reading and language traits (Paracchini et al., 2011; Scerri et al., 2011; Newbury et al., 2011).

ROBO1 (3p12)

Although a relatively small number of associations has been reported for *ROBO1* (roundabout homolog 1), this gene is considered one of the most convincing candidates implicated in reading and language skills. *ROBO1* codes for an axonal guidance receptor, which drives dendrites in the brain and has a role in several neurodevelopmental processes, including neuronal migration, branching and axonal crossing on the left-right axis in the brain (Seeger et al., 1993; Kidd et al., 1998; Andrews et al., 2006; 2008). It was first found to be disrupted by a translocation t(3;8)(p12;q11) in a dyslexic subject (Hannoula-Jouppi et al., 2005). In the same study, a rare *ROBO1* haplotype, reducing the expression of the gene, co-segregated with RD in 19 out of 21 dyslexic subjects, in a large four-generation Finnish family where linkage to 3p12 region had been originally reported (Nopola-Hemmi et al., 2001). Nonetheless, three SNPs belonging to this haplotype (6227C>A, 6483T>A and 6923T>G) and located in the 3' untranslated region (UTR) of the gene showed no association with RD status in a case-control study on 157 Indian dyslexics and 212 controls (Venkatesh et al., 2013). More recently, interesting associations were detected in a study analyzing different reading and language-related traits in an Australian population cohort (N ~1,100;

Bates et al., 2011). Among these, associations of rs6803202 and rs4535189 with nonword repetition survived correction for multiple testing, leading to the hypothesis that some common genetic variants in this gene may contribute to interindividual variation in reading and language skills in the general population (Bates et al., 2011). A replication of these associations was recently attempted by Tran and colleagues (2014) in a family-based association analysis of two RD Canadian datasets (N ~600 and ~700, respectively). Although no replication was reported, two SNPs were significantly associated with RD, namely rs331142 and rs12495133, located within a putative enhancer affecting *ROBO1* expression in the frontal cortex. These SNPs were also associated with quantitative reading and spelling traits, although these latter associations did not withstand correction for multiple testing (Tran et al., 2014).

FOXP2 (7q31.1)

FOXP2 (Forkhead box P2) was originally discovered in a multi-generational family affected by a severe monogenic form of speech and language delay, Childhood Apraxia of Speech (CAS; Lai et al., 2001). Half of the people in this family presented with difficulties in the articulation of oral speech, often accompanied by oral and written language deficits (Fisher & Scharff, 2009). A private missense mutation in this gene (R553H), determining an arginine-to-histidine substitution in the DNA-binding domain of the encoded protein, was found to be present in all the affected individuals and considered to be the disruptive variant. In an independent set of unrelated cases with a diagnosis of CAS, a nonsense mutation was found in a proband (R328X), determining an early truncation of the encoded protein (MacDermot et al., 2005). This mutation co-segregated with the disorder in the family of the proband, further supporting the involvement of the gene in language abilities. Additional evidence came from the detection of different structural aberrations, such as translocations and deletions, disrupting *FOXP2* in subjects affected with CAS (see Fisher and Scharff, 2009; Graham and Fisher, 2013 for a review). In light of these studies demonstrating the involvement of *FOXP2* in language abilities, candidate SNP association analyses tried to detect potential links with complex language disorders. Rice at al. (2009) reported association for two SNPs, rs17137124 and rs12705970, with an omnibus language measure and a measure of speech articulation, in a single set of 86 families ascertained for SLI (N=322). However, association p-values were only nominally significant and no replication of these associations was provided. Peter et al. (2011) later assessed 5 distinct *FOXP2* SNPs for association with both

RD and several reading-related traits in 188 family trios with dyslexia. This study reported several nominally significant associations, but only the association between rs7782412 and a timed word reading measure survived Bonferroni correction (Peter et al., 2011). The same SNP was also nominally associated with nonword repetition and a measure of motor sequencing. More recently, Wilcke et al. (2012) reported nominally significant association of rs12533005 with RD in a case-control association analysis of 61 dyslexics and 184 normal readers. *FOXP2* encodes a transcription factor which is thought to regulate several biological functions. In the CNS, these include signal transduction, neurite outgrowth, axon guidance, neurotransmission and synaptic plasticity (Fisher and Scharff, 2009).

CNTNAP2 (7q35-q36.1)

CNTNAP2 (contactin-associated protein-like 2) is a molecular target of FOXP2, and has been implicated in language-related phenotypes in multiple studies. This gene encodes CASPR2, a protein facilitating cell-cell interaction and adhesion in the CNS (Rodenas-Cuadrado et al., 2014). This protein is highly expressed in the developing brain, where it is thought to have a fundamental role in neuronal migration, dendrite outgrowth and clustering of voltage-gated ion channels at Ranvier nodes (Rodenas-Cuadrado et al., 2014). Different polymorphisms/mutations/structural aberrations in *CNTNAP2* have been implicated in several neurodevelopmental and neuropsychiatric disorders, including Autism Spectrum Disorders (ASD), epilepsy, Attention Deficit Hyperactivity Disorder (ADHD) and other learning disabilities (reviewed in Newbury & Monaco, 2010; Rodenas-Cuadrado et al., 2014). The first evidence of involvement of this gene in SLI etiology came from a study testing association between several SNPs in *CNTNAP2* and three continuous language traits, namely nonword repetition, expressive and receptive language score (Vernes et al., 2008). This targeted association analysis -which involved 184 SLI families from the SLIC dataset used in this study- led to detection of nine significant associations within introns 13-14, both at the single SNP and at the haplotype level. The most prominent associations were observed with nonword repetition, at rs10246256, rs17236239, rs2710117 and rs2710102 (Vernes et al., 2008). These SNPs were later tested in a similar sample of SLIC families, in the context of a wider assessment of several candidate SNPs from different genes previously implicated in RD and/or SLI (Newbury et al., 2011). This work confirmed the associations reported by Vernes and colleagues (2008), and detected novel associations of these SNPs with reading-related skills, such as word reading, spelling and comprehension (Newbury et al., 2011).

These findings suggested pleiotropic effects of these variants on several reading and language traits in SLI populations, a hypothesis further corroborated by the nominally significant association found between rs2710102 and nonword repetition in a sample of dyslexic family trios (Peter et al., 2011). The same region (between exons 13 and 15) showed an effect on language skills also in the general population, as suggested by a candidate SNP association analysis of early communicative behavior in ~1,150 two-years-old children from an Australian cohort (Whitehouse et al., 2011). SNP-based analysis reported nominally significant associations for rs2710102 and rs759178; while haplotype analysis revealed a significant association surviving Bonferroni correction for the haplotype rs2710102–rs759178–rs17236239–rs2538976, overlapping with the nine-markers haplotype associated with nonword repetition in Vernes et al. (2008). Further support for the effect of this region on language skills comes from the suggestive associations reported for rs2710102 and rs1718101 with two language endophenotypes of ASD, namely age at first word (Alarcon et al., 2008) and age at first phrase (Anney et al., 2012). Nonetheless, Toma et al. (2013) did not find any evidence of association of rs2710102 with these two endophenotypes.

CMIP (16q23.2-q23.3)

A targeted association analysis in the SLIC dataset (806 individuals from 211 families) investigated a known candidate region previously linked to SLI (*SLI1*) and led to the discovery of two genes significantly associated with nonword repetition performance, *CMIP* and *ATP2C2* (Newbury et al., 2009). *CMIP* (c-MAF induced protein) encodes an adaptor protein which may act as a cytoskeletal component. This suggests that also C-MIP may contribute to the neuron migration process, a hypothesis supported by its interaction with the neuronal migration protein filamin A (Grimbert et al., 2004). In *CMIP*, Newbury and colleagues (2009) observed associations at several SNPs between exons 2 and 5, with top association at rs6564903. These associations were detected in SLIC both in quantitative trait analysis and in case-control analysis, and were internally replicated in a language-impaired sample (112 cases) selected from a British population-based cohort (490 subjects), although with an opposite direction of effect. However, no association was observed with continuous nonword repetition trait in the whole unselected cohort (Newbury et al., 2009). Some of these SNPs later showed nominally significant associations with word reading and spelling, both in the SLIC sample (for rs6564903, rs12927866, rs7201632 and rs3935802; Newbury et al., 2011), and in a general population cohort of ~3,700 subjects (for rs6564903, rs12927866 and

rs16955705; Scerri et al., 2011), suggesting pleiotropic effects of *CMIP* on reading and language skills.

ATP2C2 (16q24.1)

Newbury and colleagues (2009) also detected significant associations independent of *CMIP* in another gene, *ATP2C2* (ATPase, Ca^{2+} transporting, type 2C, member 2). This gene codes for a calcium transporter ATPase regulating cellular levels of calcium and manganese ions, a key process for synaptic plasticity, transmission and neuronal motility (Newbury & Monaco; 2010). Significant associations in this gene were detected between exons 7 and 12 (top hit rs11860694), and were internally replicated in the set of SLI cases selected from a population-based cohort, with consistent allelic trends (Newbury et al., 2009). Moreover, these SNPs showed borderline significant associations with a composite measure of receptive and expressive language in SLIC (Newbury et al., 2009; 2011). Associations of *ATP2C2*, however, did not extend to RD-related traits such as word reading and spelling, as suggested by the lack of associations reported in two independent studies (Newbury et al., 2011; Scerri et al., 2011).

The present study

In the present chapter, we test genetic associations in the candidate genes reviewed above, first with a principal component score derived from several reading and language traits and then with the individual reading and language traits available within each of the datasets previously involved in our GWAS meta-analysis (described in Chapter 3). Both candidate SNP and candidate gene-based association analyses were carried out. Since most of the SNPs examined in this work were previously identified or analysed in studies partially overlapping with our datasets (see below), the main aim of this chapter was not to provide independent replications of these findings, but rather to evaluate the consistency with original findings and the patterns of pleiotropic associations for these candidate SNPs. Nonetheless, some of the SNP associations assessed -namely those in *ROBO1* and *FOXP2*- were originally reported in datasets other than those analysed here and were never replicated. Therefore, for these SNPs we attempted to detect independent support of previous findings.

Several studies have attempted to replicate previous associations with RD/SLI (reviewed above). The results of our assessment will be compared with these studies and will be discussed to draw general conclusions on the replicability of these findings.

Subjects and Methods

We assessed associations of candidate SNPs and genes (see *Candidate SNPs and genes analysed* below) with a principal component score extracted from several reading and language traits, in a meta-analysis of >1,800 subjects from three datasets. A detailed description of the datasets and of the methods used can be found in Chapter 2 and 3. We provide below a brief summary, which is essential to the understanding of the present chapter.

Datasets, phenotypic measures and genetic association analyses with PC traits

Three datasets of children affected by reading and language deficits and their siblings were involved in the study: the UK-RD dataset (N=983, mean age 11.7 years, age range 5-31), comprising children diagnosed with RD and their siblings (608 nuclear families), from United Kingdom; the SLIC dataset (N=548, mean age 10 years, age range 5-19), which comprised children affected with SLI and their siblings (288 nuclear families), from UK; and the CLDRC dataset (N=749, mean age 11.7 years, age range 8-19), composed by twins recruited in Colorado (US) for a school history of RD or ADHD, along with their co-siblings (343 unrelated twinships/sibships). Of these, 266 twinships/sibships (N=585) were recruited via a proband with a history of RD, and 77 twinships/sibships (N=164) were recruited via a proband with a history of ADHD. These subsets, named CLDRC-RD and CLDRC-ADHD respectively, were analyzed separately. In each dataset, participants had been assessed for a number of reading and language abilities through psychometric tests, regardless of diagnosis. The traits available in UK-RD were word reading (WRead), spelling (WSpell), phonological decoding (PD), phoneme awareness (PA) and orthographic coding (OC). In SLIC they were WRead, WSpell, nonword repetition (NWR), expressive (ELS) and receptive language scores (RLS). In both CLDRC datasets (CLDRC-RD and CLDRC-ADHD), the abilities assessed were WRead, WSpell, PD, PA, OC and NWR. Reading and language traits had been previously age-adjusted according to normative data, and underwent a further rank-normalization when required, to attain normality of distributions within datasets. Phenotypic outliers for three or more trait scores and subjects with full scale IQ < 70 were discarded.

Genome-wide genotype data underwent a first round of quality control in PLINK v1.07 (Purcell et al., 2007). All SNPs deviating from Hardy-Weinberg Equilibrium (HWE, $p < 1x10^{-6}$), with Minor Allele Frequency (MAF) < 1%, and call frequency < 99%, were filtered

out. Similarly, samples were excluded if they showed inconsistencies in genome-wide identity-by-descent sharing with their siblings and unrelated individuals, or sex mismatches, or call rates < 98%. Outliers in Multi-Dimensional Scaling (MDS) analysis of genome-wide genotype data and outliers for genome-wide homozygosity were also discarded. Then genotype data underwent phasing in MACH v1.0 (Li et al., 2010) and imputation on the 1000 Genomes Project reference (GIANT all populations panel, Phase 1, v3; The 1000 Genomes Project Consortium, 2012) using Minimac (Howie et al., 2012). A final quality control procedure was then run on the imputed data, first excluding poorly imputed polymorphisms (with $r^2 < 0.3$) and individual genotypes with imputation quality scores < 0.9, and then discarding SNPs with HWE $p < 5x10^{-6}$, MAF < 1%, and call frequency < 95%.

We extracted the first principal component (PC1) within each dataset from all the reading and language traits available. PC1 explained a substantial proportion (52-75%) of common variance in reading and language traits in the datasets. In addition, we residualized PC1 against performance IQ to compute an IQ-adjusted PC1 score, which was also assessed. At the end of these processes, we had PC1 data for 544 participants in CLDRC-RD, 914 participants in UK-RD, 245 participants in SLIC and 159 participants in CLDRC-ADHD, for a total of 1,862 subjects involved in the PC1 meta-analysis. Sample sizes of IQ-adjusted PC1 analyses were N=544 in CLDRC-RD, N=878 in UK-RD, N=245 in SLIC and N=159 in CLDRC-ADHD, for a final sample size of 1,826 in the IQ-adjusted PC1 meta-analysis.

Sibling-based association analyses of PC1 and IQ-adjusted PC1 were run as described in Chapter 3. First, associations were analysed separately within each dataset, through PLINK QFAM 'total' association analysis, and then they were meta-analysed together through the *Sample Size Scheme* in METAL (Willer et al., 2010). Finally, gene-based association analyses were carried out using VEGAS v0.8.27 (Liu et al. 2010). Each tested gene also included potentially regulatory regions located up to 50 kb beyond the 5'- and 3'-untranslated regions (UTRs).

Candidate SNPs and genes analysed

In our assessment, we included variants within nine candidate loci previously implicated in RD/SLI, namely *DYX1C1, KIAA0319, DCDC2, MRPL19/GCFC2, ROBO1, CNTNAP2, CMIP, ATP2C2* and *FOXP2* (see *Introduction*).

We adopted a staged approach to study the candidates. We initially aimed to test 30 candidate SNPs which had been previously reported to be associated with RD/SLI and related traits in two or more independent studies/datasets, either directly (at the same SNP) or indirectly (in the same haplotype), as reviewed in the *Introduction* section. These SNPs were rs1000585, rs917235, rs714939, and rs6732511 in *MRPL19/GCFC2*; rs793862, rs807701, rs807724, rs1087266 in *DCDC2*; rs4504469, rs761100, rs6935076, rs3212236, rs9461045 and rs2143340 in *KIAA0319*; rs57809907 and rs3743205 in *DYX1C1*; rs10246256, rs2710102, rs17236239 and rs2710117 in *CNTNAP2*; rs12927866, rs6564903, rs3935802, rs4265801, rs7201632 in *CMIP*; and rs8053211, rs11860694, rs16973771, rs2875891 and rs8045507 in *ATP2C2*. These SNPs have already been analysed in a candidate gene association study to test their effect on several reading and language traits (Newbury et al., 2011). Here, we decided to extend this analysis, testing additional datasets (i.e. CLDRC-RD and CLDRC-ADHD) and additional traits (i.e. PC1 and IQ-adjusted PC1). Of these 30 SNPs, only 25 were available in our genotype data (see Table 1a, b); the remaining five had been poorly imputed or discarded in the QC of one or more of the datasets.

We also investigated an additional set of five SNPS, namely rs6803202, rs4535189, rs331142 and rs12495133 in *ROBO1*, and rs7782412 in *FOXP2*, which were not included in the analysis by Newbury and colleagues (2011). These SNPs have been associated with reading and language phenotypes by single studies on datasets other than those involved in our analysis (Bates et al., 2011; Peter et al., 2011; Tran et al., 2014), but their associations have been never replicated.

For the analysis of the 30 candidate SNPs mentioned above, we did not to use any correction for multiple testing, as for these associations the main aim was to evaluate their consistency with original findings and their cross-phenotypic effects (see below).

To investigate the presence of other association signals in the nine candidate RD/SLI genes tested, we extended our SNP assessment to all the SNPs falling within coding sequences and in their potential regulatory regions, up to 50 kb beyond the 5'- and 3'-UTRs. Following these criteria, a total of 13,827 SNPs were assessed in these candidate genes: 735 in *MRPL19/GCFC2*; 2,261 in *ROBO1*; 899 in *DCDC2*; 717 in *KIAA0319*; 453 in *FOXP2*; 6,473 in *CNTNAP2*; 520 in *DYX1C1*; 827 in *CMIP* and 942 in *ATP2C2*.

In this gene-wide analysis, we applied an appropriate Bonferroni correction for the number of traits and SNPs tested. We first accounted for the interdependence of SNPs tested by

calculating the number of independent tests in our candidate genes, through the *Genetic Type I error calculator* method (Li et al., 2012; http://statgenpro.psychiatry.hku.hk/gec/index.php), applied to each of our datasets. We took the highest number of independent tests computed in our datasets, namely 2,130 in UK-RD (versus 2,074 in CLDRC and 2,042 in SLIC). Then we corrected the significance (α) threshold for this number, and finally for the number of traits tested (2, i.e. PC1 and IQ-adjusted PC1). This resulted in a final corrected α threshold of $1.2x10^{-5}$.

We also assessed the gene-based associations of our candidate genes (representing the combined associations of all the SNPs within those gene, adjusted for their LD structure), and investigated the regional association patterns with both PC1 and IQ-adjusted PC1, by plotting all the SNPs showing association $p < 0.1$, through LocusZoom (http://csg.sph.umich.edu/locuszoom/; Pruim et al., 2010).

Association analysis of candidate SNPs with individual reading/language traits

For five candidate SNPs which were found to be associated with PC1/IQ-adjusted PC1 (see *Results* section below), here we further investigated the pattern of cross-phenotypic associations in each dataset. We first ran QFAM univariate association analysis in PLINK v1.07 (Purcell et al., 2007), for each individual trait used in PC1 computation. Then we performed multivariate association analysis through PLINK Multivariate v1.06 (Ferreira & Purcell, 2009), on the same set of reading/language traits that were used in constructing PC1. PLINK Multivariate extracts the linear combination of traits that explains the largest possible amount of covariance between the SNP and the traits analysed. The loading produced for each trait represents its contribution to the multivariate association. MQFAM 'total' association was run, with adaptive permutations to adjust for sample relatedness (see *Supplementary Material S1* in Chapter 3 for further details on this analysis).

Results

Below we report association p-values uncorrected for multiple testing, indicating the corrected α threshold used only where applicable (i.e. in the gene-wide SNP analysis, see *Subjects and Methods* section and further below).

Assessment of SNPs and genes implicated in RD and/or SLI

Initially we assessed 25 SNPs, for which evidence of association with reading and language traits has been reported in two or more independent studies/datasets, either directly -at the same SNP- or indirectly -in the same haplotype (see *Introduction* and Carrion-Castillo et al., 2013; Newbury et al., 2010; Rodenas-Cuadrado et al., 2014). The results of this assessment are reported in Table 1a, b. The most significant association was with SNPs within or close to *KIAA0319*, most notably for rs2143340 (PC1 p = $1.4x10^{-4}$; IQ-adjusted PC1 p = $8.3x10^{-4}$; A/G, minor allele G, MAF ~15%), rs3212236 (PC1 p = $2x10^{-3}$; IQ-adjusted PC1 p =$8.5x10^{-3}$; T/C, minor allele C; MAF ~17%), rs9461045 (PC1 p = $2.7x10^{-3}$; IQ-adjusted PC1 p = $9.6x10^{-3}$; T/C, minor allele T; MAF ~17%), and rs761100 (PC1 p = 0.02; IQ-adjusted PC1 p = 0.03; A/C, minor allele C; MAF ~45%). Three SNPs in *ATP2C2*, namely rs16973771 (T/C, minor allele C, MAF ~41%), rs2875891 (T/C, minor allele T, MAF ~36%) and rs8045507 (A/G, minor allele A, MAF ~40%), also showed nominally significant evidence for association (p~ 0.029-0.045, see Table 1a, b). For all these SNPs, the allelic trends were concordant with those reported in the original studies (see Table 1a, b), as reviewed by Newbury et al. (2011).

Then we assessed five additional SNPs (4 in *ROBO1* and 1 in *FOXP2*) for which significant associations had been reported in the past (Bates et al., 2011; Peter et al., 2011; Tran et al., 2014; Wilcke et al., 2012), but were never replicated. The results of this assessment are shown in Table 2a, b. We observed association p-values < 0.05 for three *ROBO1* polymorphisms -rs6803202 (T/C, minor allele C, MAF ~49.6%), rs4535189 (A/G, minor allele A, MAF ~49.5%) and rs12495133 (A/C, minor allele A, MAF ~36%)- with both PC1 and IQ-adjusted PC1. Of these SNPs, rs12495133 was the most significantly associated with our traits of interest (p = $7.9x10^{-4}$ with PC1 and p = $4.1x10^{-4}$ with IQ-adjusted PC1). The two largest datasets in our meta-analysis, CLDRC-RD and UK-RD, provided the strongest contribution to this association, as revealed by the weighted Z-scores of rs12495133 association in each dataset as computed by METAL (Table S1a). Nonetheless, effect sizes were comparable across all the datasets (see QFAM beta values in Table S1a). The allelic trend was consistent with the original report (Tran et al., 2014), with minor allele A showing a positive effect on reading and language skills. Conversely, associations at rs6803202 and rs4535189 (p ~ 0.02-0.03 with PC1 and ~ $8x10^{-3}$ with IQ-adjusted PC1) showed allelic trends opposite to those detected in Bates et al. (2011), with allele T of rs6803202 and allele G of rs4535189 exerting a negative effect on our traits of interest (see Table 2a, b).

Finally, we extended the SNP assessment at the gene-wide level, in all nine of the candidate genes tested (see *Subjects and Methods* section). The results of this assessment (reported in Table S1b, c) revealed no significant association withstanding correction for multiple testing of two traits (PC1 and IQ-adjusted PC1) and 2,130 independent SNPs tested ($\alpha = 1.2x10^{-5}$; see *Subjects and Methods*). The most significant association was detected at rs3181234 (*KIAA0319*; $p \sim 5.2 \times 10^{-5}$, T/C, minor allele T, MAF ~15%) with PC1, and at rs2311350 (*ROBO1*; $p \sim 2.9 \times 10^{-5}$, A/G, minor allele G, MAF ~33%) with IQ-adjusted PC1. These SNPs were among the top associations ($p < 0.001$) with both PC1 and IQ-adjusted PC1, and the minor alleles showed negative allelic trends with these traits. The majority of the most significant associations observed with PC1 and IQ-adjusted PC1 (Table S1b, c) were annotated to *KIAA0319* and *ROBO1*. Regional association plots of *KIAA0319* (Figure 1) and *ROBO1* (Figure 2) showed the presence of distinct association signals in low LD in these genes. In addition to these two genes, only *DCDC2* showed two SNPs among the top associations with PC1 (rs146260219 and rs114966185, $p \sim 5x10^{-4}$), but not with IQ-adjusted PC1 ($p > 0.001$).

In line with the results of SNP-based meta-analysis, gene-based association tests (Table 3a, b) revealed associations of *ROBO1* and *KIAA0319*, both with PC1 ($p \sim 5x10^{-3}$ and $\sim 6x10^{-3}$) and with IQ-adjusted PC1 ($p \sim 1x10^{-3}$ and ~ 0.03). All the other genes showed no evidence of association in the gene-based analysis.

1a)

Chr	Gene	SNP	Position	Allele1	Allele2	Freq (%)[b]	P-value	Direction[c]	Original Report	Phenotype
2	MRPL19/GCFC2	rs1000585	75823162	a	g	60.02	0.141	++-+	Anthoni et al. 2007	RD
2	MRPL19/GCFC2	rs917235	75825819	a	g	52.11	0.28	++-+	Anthoni et al. 2007	RD
2	MRPL19/GCFC2	rs714939	75835107	a	g	39.69	0.894	+--+	Anthoni et al. 2007	RD
2	MRPL19/GCFC2	rs6732511	75839733	t	c	16.96	0.849	-++-	Anthoni et al. 2007	RD
6	DCDC2	rs793862	24207200	a	g	26.7	0.999	++--	Schumacher et al. 2006	RD
6	DCDC2	rs807701	24273791	a	g	64.63	0.913	-++-	Schumacher et al. 2006	RD
6	DCDC2	rs807724	24278869	t	c	78.07	0.828	+-+-	Meng et al. 2005	RD
6	**KIAA0319**	**rs761100**	**24632642**	**a**	**c**	**45.04**	**0.02**	**++-+**	**Harold et al. 2006**	**RD**
6	KIAA0319	rs6935076	24644322	t	c	36.13	0.909	--++	Cope et al. 2005	RD
6	**KIAA0319**	**rs3212236[a]**	**24648455**	**t**	**c**	**83.24**	**2×10^{-3}**	**++++**	**Harold et al., 2006**	**RD**
6	**KIAA0319**	**rs9461045**	**24649061**	**t**	**c**	**16.79**	**2.7×10^{-3}**	**----**	**Dennis et al. 2009**	**RD**
6	**KIAA0319**	**rs2143340**	**24659071**	**a**	**g**	**85.34**	**1.4×10^{-4}**	**++++**	**Francks et al. 2004**	**RD**
15	DYX1C1	rs57809907[a]	55722882	a	c	8.52	0.907	-+--	Taipale et al. 2003	RD
15	DYX1C1	rs3743205[a]	55790530	t	c	5.9	0.894	-++-	Taipale et al. 2003	RD
7	CNTNAP2	rs10246256	147554807	t	c	68.82	0.32	-+-+	Vernes et al. 2008	SLI
7	CNTNAP2	rs2710102	147574390	a	g	50.55	0.839	--+-	Vernes et al. 2008	SLI
7	CNTNAP2	rs17236239	147582305	a	g	65.58	0.373	-+++	Vernes et al. 2008	SLI
7	CNTNAP2	rs2710117	147601772	a	t	64.47	0.708	-+-+	Vernes et al. 2008	SLI
16	CMIP	rs12927866	81652322	t	c	40.2	0.812	--+-	Newbury et al. 2009	SLI
16	CMIP	rs3935802	81661567	c	g	41.44	0.869	--+-	Newbury et al. 2009	SLI
16	ATP2C2	rs8053211	84453753	a	g	52.92	0.199	----	Newbury et al. 2009	SLI
16	ATP2C2	rs11860694	84457447	c	g	46.81	0.074	++++	Newbury et al. 2009	SLI
16	**ATP2C2**	**rs16973771**	**84460578**	**t**	**c**	**59.47**	**0.036**	**----**	**Newbury et al. 2009**	**SLI**
16	**ATP2C2**	**rs2875891**	**84463909**	**t**	**c**	**35.78**	**0.029**	**+++-**	**Newbury et al. 2009**	**SLI**
16	**ATP2C2**	**rs8045507**	**84464577**	**a**	**g**	**40.27**	**0.041**	**++++**	**Newbury et al. 2009**	**SLI**

1b)

Chr	Gene	SNP	Position	Allele1	Allele2	Freq (%)[b]	P-value	Direction[c]	Original Report	Phenotype
2	MRPL19/GCFC2	rs1000585	75823162	a	g	60.02	0.129	++-+	Anthoni et al. 2007	RD
2	MRPL19/GCFC2	rs917235	75825819	a	g	52.11	0.24	++-+	Anthoni et al. 2007	RD
2	MRPL19/GCFC2	rs714939	75835107	a	g	39.69	0.601	+--+	Anthoni et al. 2007	RD
2	MRPL19/GCFC2	rs6732511	75839733	t	c	16.96	0.51	-++-	Anthoni et al. 2007	RD
6	DCDC2	rs793862	24207200	a	g	26.7	0.611	++--	Schumacher et al. 2006	RD
6	DCDC2	rs807701	24273791	a	g	64.63	0.736	-++-	Schumacher et al. 2006	RD
6	DCDC2	rs807724	24278869	t	c	78.07	0.584	--++	Meng et al. 2005	RD
6	**KIAA0319**	**rs761100**	**24632642**	**a**	**c**	**45.04**	**0.03**	**++-+**	**Harold et al. 2006**	**RD**
6	KIAA0319	rs6935076	24644322	t	c	36.13	0.625	--+-	Cope et al. 2005	RD
6	**KIAA0319**	**rs3212236[a]**	**24648455**	**t**	**c**	**83.24**	**8.5×10^{-3}**	**++++**	**Harold et al., 2006**	**RD**
6	**KIAA0319**	**rs9461045**	**24649061**	**t**	**c**	**16.79**	**9.6×10^{-3}**	**----**	**Dennis et al. 2009**	**RD**
6	**KIAA0319**	**rs2143340**	**24659071**	**a**	**g**	**85.34**	**8.4×10^{-4}**	**++++**	**Francks et al. 2004**	**RD**
15	DYX1C1	rs57809907[a]	55722882	a	c	8.52	0.573	+++-	Taipale et al. 2003	RD
15	DYX1C1	rs3743205[a]	55790530	t	c	5.9	0.488	-++-	Taipale et al. 2003	RD
7	CNTNAP2	rs10246256	147554807	t	c	68.82	0.288	---+	Vernes et al. 2008	SLI
7	CNTNAP2	rs2710102	147574390	a	g	50.55	0.989	+-+-	Vernes et al. 2008	SLI
7	CNTNAP2	rs17236239	147582305	a	g	65.58	0.25	+++-	Vernes et al. 2008	SLI
7	CNTNAP2	rs2710117	147601772	a	t	64.47	0.804	-+-+	Vernes et al. 2008	SLI
16	CMIP	rs12927866	81652322	t	c	40.2	0.854	--+-	Newbury et al. 2009	SLI
16	CMIP	rs3935802	81661567	c	g	41.44	0.944	--+-	Newbury et al. 2009	SLI
16	ATP2C2	rs8053211	84453753	a	g	52.92	0.15	---+	Newbury et al. 2009	SLI
16	ATP2C2	rs11860694	84457447	c	g	46.81	0.059	+++-	Newbury et al. 2009	SLI
16	**ATP2C2**	**rs16973771**	**84460578**	**t**	**c**	**59.47**	**0.042**	**----**	**Newbury et al. 2009**	**SLI**
16	**ATP2C2**	**rs2875891**	**84463909**	**t**	**c**	**35.78**	**0.037**	**+++-**	**Newbury et al. 2009**	**SLI**
16	**ATP2C2**	**rs8045507**	**84464577**	**a**	**g**	**40.27**	**0.045**	**++++**	**Newbury et al. 2009**	**SLI**

Table 1. Assessment of candidate SNPs previously associated with RD/SLI and related traits in two or more independent studies/datasets. Here association p-values with **a)** PC1 and **b)** IQ-adjusted PC1 are shown. Nominally significant associations ($p < 0.05$) are highlighted in bold. For none of these associated SNPs a significant heterogeneity of effect size was detected across the different datasets (heterogeneity $p > 0.05$).[a] These SNPs are reported in the opposite strand compared to the original studies. [b] Frequency of Allele 1. [c] The direction of effect of Allele 1 is reported for datasets in the following order: CLDRC-RD, UK-RD, SLIC, CLDRC-ADHD.

2a)

Chr	Gene	SNP	Position	Allele1	Allele2	Freq (%)[a]	P-value	Direction[b]	Original Report	Phenotype[c]
3	**ROBO1**	**rs6803202**	**79499153**	**t**	**c**	**50.43**	**0.027**	**---+**	**Bates et al (2011)**	**NWR**
3	**ROBO1**	**rs4535189**	**79489971**	**a**	**g**	**49.54**	**0.021**	**+++-**	**Bates et al (2011)**	**NWR**
3	ROBO1	rs331142	78920844	a	c	74.98	0.438	+-++	Tran et al. (2014)	RD
3	**ROBO1**	**rs12495133**	**78921520**	**a**	**c**	**36.34**	**7.9×10^{-4}**	**++++**	**Tran et al. (2014)**	**RD**
7	FOXP2	rs7782412	114290415	t	c	56.38	0.528	++-+	Peter et al. (2011)	WRead

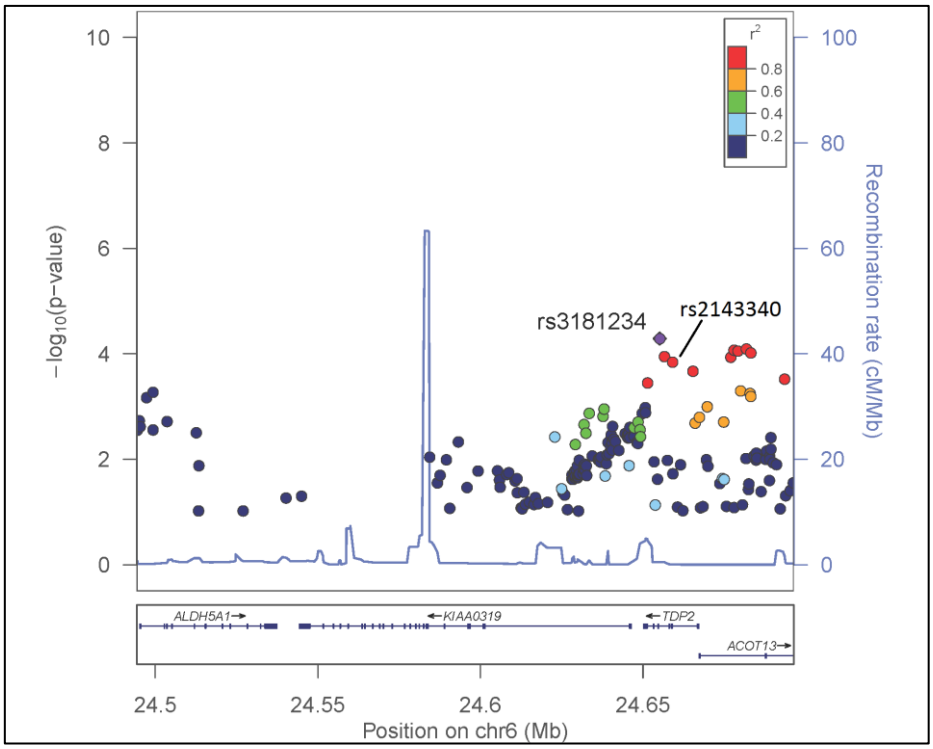
2b)

Chr	Gene	SNP	Position	Allele1	Allele2	Freq (%)[a]	P-value	Direction[b]	Original Report	Phenotype[c]
3	**ROBO1**	**rs6803202**	**79499153**	**t**	**c**	**50.43**	**8.4×10^{-3}**	**---+**	**Bates et al (2011)**	**NWR**
3	**ROBO1**	**rs4535189**	**79489971**	**a**	**g**	**49.54**	**8.4×10^{-3}**	**+++-**	**Bates et al (2011)**	**NWR**
3	ROBO1	rs331142	78920844	a	c	74.98	0.493	+-++	Tran et al. (2014)	RD
3	**ROBO1**	**rs12495133**	**78921520**	**a**	**c**	**36.34**	**4.1×10^{-4}**	**++++**	**Tran et al. (2014)**	**RD**
7	FOXP2	rs7782412	114290415	t	c	56.38	0.474	++-+	Peter et al. (2011)	WRead

Table 2. Assessment of five additional candidate SNPs lying within *ROBO1* and *FOXP2*. These SNPs have been associated with reading and language phenotypes by single studies (see table), but their associations have been never replicated. Here, association p-values with **a)** PC1 and **b)** IQ-adjusted PC1 are shown. Nominally significant associations ($p < 0.05$) are highlighted in bold. For none of these associated SNPs a significant heterogeneity of effect size was detected across the different datasets (heterogeneity $p > 0.05$).

[a] Frequency of Allele 1. [b] The direction of effect of Allele 1 is reported for datasets in the following order: CLDRC-RD, UK-RD, SLIC, CLDRC-ADHD.[c] Phenotype associated in the original report.

a)



b)

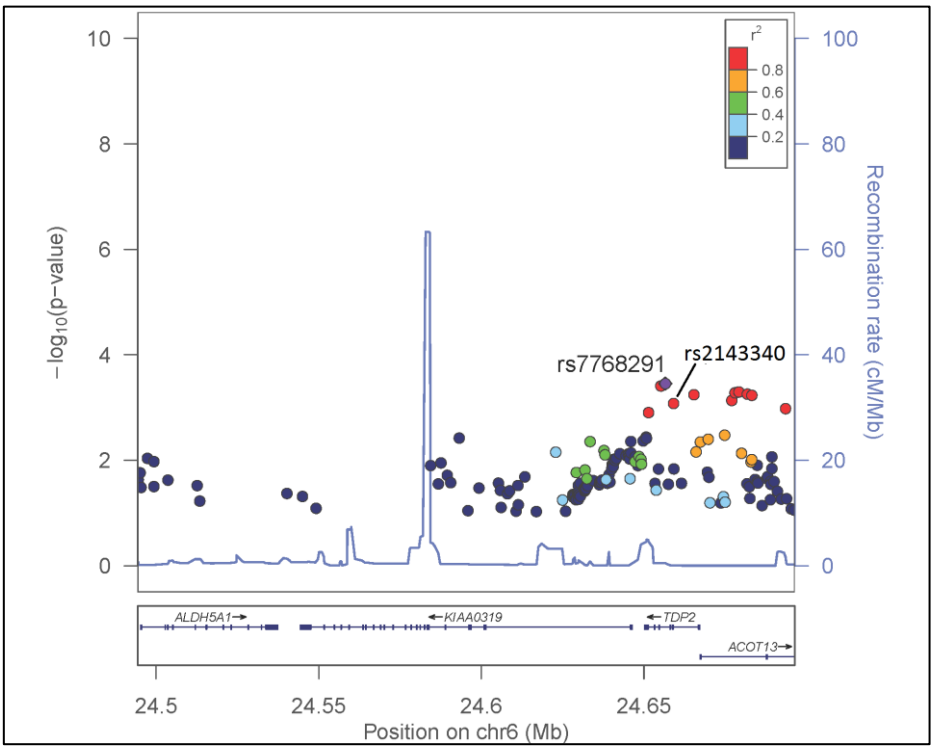
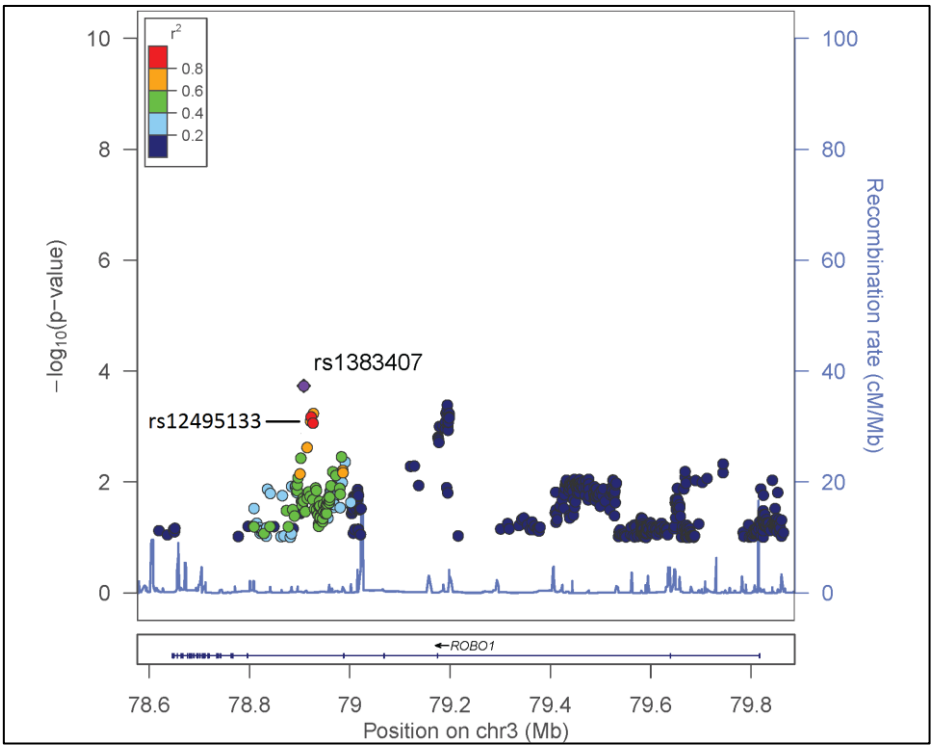


Figure 1. Regional association plot of *KIAA0319* (6p22). All the SNPs showing association p < 0.1 with **a)** PC1 and **b)** IQ-adjusted PC1 are plotted. The local top hit is highlighted in violet. The candidate SNP most significantly associated in the region (rs2143340) is also indicated.

a)



b)

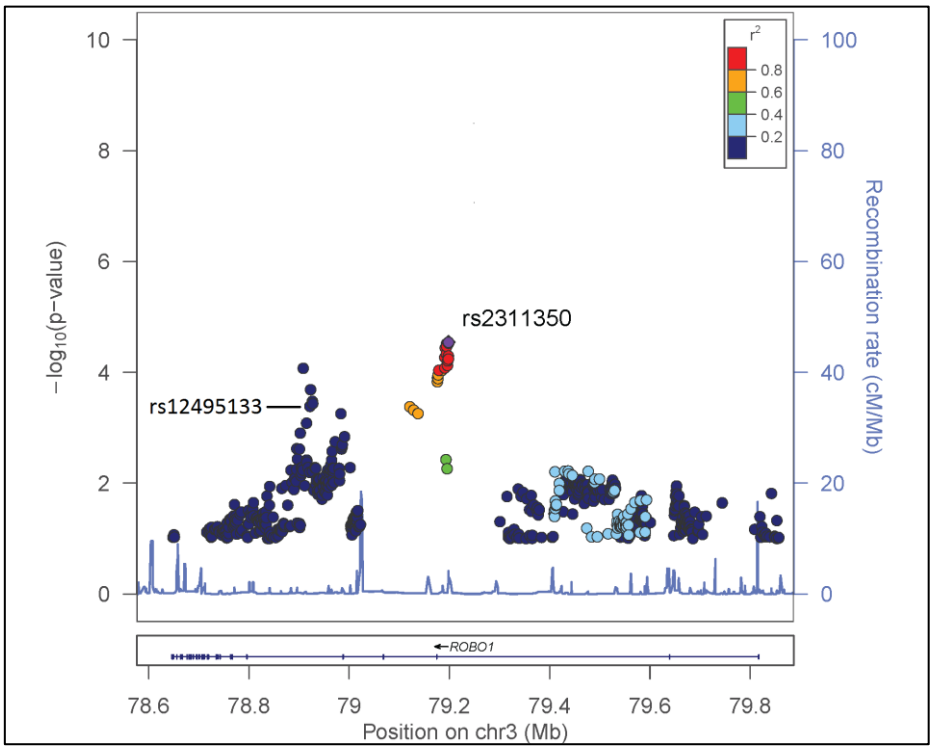


Figure 2. Regional association plot of *ROBO1* (3p12). All the SNPs showing association p < 0.1 with **a)** PC1 and **b)** IQ-adjusted PC1 are shown. The local top hit is highlighted in violet. The candidate SNP most significantly associated in the region (rs12495133) is also indicated.

3a)

Chr	Gene	nSNPs[a]	Start[b]	Stop[b]	Pvalue	Best-SNP[a]	SNP-pvalue
6	**KIAA0319**	**260**	**24652310**	**24754362**	**6.2×10^{-3}**	**rs7768291**	**1.1×10^{-4}**
6	DCDC2	284	24279961	24466259	0.992	rs16889066	0.074
3	**ROBO1**	**589**	**78729079**	**79721751**	**5×10^{-3}**	**rs1383407**	**1.9×10^{-4}**
15	DYX1C1	102	53497245	53587724	0.825	rs6493791	0.099
2	MRPL19	133	75727416	75742842	0.566	rs17690622	0.021
2	GCFC2	142	75742801	75791830	0.586	rs17690622	0.021
7	FOXP2	222	113842287	114118328	0.594	rs11762537	0.036
7	CNTNAP2	2708	145444385	147749019	0.827	rs10257633	1×10^{-3}
16	CMIP	344	80036275	80302868	0.715	rs4243209	0.01
16	ATP2C2	338	82959633	83055294	0.385	rs173429	3.4×10^{-3}

3b)

Chr	Gene	nSNPs[a]	Start[b]	Stop[b]	Pvalue	Best-SNP[a]	SNP-pvalue
6	**KIAA0319**	**260**	**24652310**	**24754362**	**0.034**	**rs7768291**	**3.5×10^{-4}**
6	DCDC2	284	24279961	24466259	0.849	rs1620407	0.011
3	**ROBO1**	**589**	**78729079**	**79721751**	**1.1×10^{-3}**	**rs2311350**	**2.9×10^{-5}**
15	DYX1C1	102	53497245	53587724	0.379	rs3759864	0.078
2	MRPL19	133	75727416	75742842	0.51	rs6547014	0.044
2	GCFC2	142	75742801	75791830	0.615	rs6547014	0.044
7	FOXP2	222	113842287	114118328	0.442	rs11762537	0.024
7	CNTNAP2	2708	145444385	147749019	0.705	rs851821	4.5×10^{-3}
16	CMIP	344	80036275	80302868	0.944	rs9972695	4.5×10^{-3}
16	ATP2C2	338	82959633	83055294	0.254	rs8055494	8.2×10^{-3}

Table 3. Gene-based (VEGAS) association tests for candidate RD/SLI genes in **a)** PC1 and **b)** IQ-adjusted PC1 meta-analysis. Genes showing nominally significant associations ($p < 0.05$) are highlighted in bold.
[a] Note that this analysis was based on SNPs included in HapMap CEU reference (release R2, The International HapMap 3 Consortium, 2010), therefore not all the polymorphisms meta-analyzed in our study (based on 1000 Genomes reference, Phase 1 V 3) were included in this test. For the same reason, the most significantly associated SNP does not necessarily coincide with other analyses based on imputation with the 1000 Genomes reference dataset. [b] Start and stop positions are expressed in hg18 coordinates, as per VEGAS output.

Cross-phenotypic effects of the most associated candidate SNPs

We further investigated the patterns of cross-phenotypic associations for five out of ten SNPs which had been found to be associated in the previous analysis of 30 candidate SNPs (Table 1, 2). These SNPs, representing the strongest associations in their regions, included rs761100 and rs2143340 in *KIAA0319*, rs2875891 in *ATP2C2*, rs4535189 and rs12495133 in *ROBO1*. When two or more associated SNPs in the same region showed low LD ($r^2 < 0.3$), suggesting that they may tag different genetic effects, we analysed both SNPs. We did so for rs761100

and rs2143340 in *KIAA0319*, and for rs4535189 and rs12495133 in *ROBO1*. The results of this association analysis with individual reading/language traits are reported in Tables S1d, e, f, g, h.

The *ROBO1* SNP rs12495133 (Table S1d) showed moderate to high multivariate association loadings with a number of traits in all the datasets, more prominently in CLDRC-RD and UK-RD. At the univariate level, nominally significant associations were observed with phonological decoding (both nonword reading and phonological choice, p = 0.017 and 0.047) in CLDRC-RD; with word reading (p = 0.039), spelling (p = 0.019), and orthographic coding (p = 0.039) in UK-RD; with receptive language score (p = 0.037) in SLIC, and with nonword repetition (p = 0.01) in CLDRC-ADHD. The direction of effect was concordant with the one shown in PC1 meta-analysis for all the traits in all the datasets, with the exception of nonword repetition in CLDRC-ADHD (Table S1d). Trends of association (with p-values falling short out of significance) were seen also with phonological decoding in UK-RD, with orthographic coding in CLDRC-RD, and with phoneme awareness in both datasets.

Another SNP in *ROBO1*, rs4535189 (Table S1e), showed nominally significant univariate associations with word spelling (p = 0.029), phonological decoding (p = 0.014), phoneme awareness (p = 0.034) and orthographic coding (p = 0.031) in UK-RD; and with expressive and receptive language scores (p = 0.006 and 0.012) in SLIC. Also in this case we observed trends of association, with nonword repetition in SLIC and with word reading in UK-RD. All these associations showed a positive effect of allele A on the individual reading and language traits, consistent with the one detected with PC scores. However, none of these associations were observed in the CLDRC datasets, where the allelic trends were discordant for most of the traits analysed (Table S1e).

For rs761100 (*KIAA0319*; Table S1f), we observed significant univariate associations in UK-RD -with word reading (p = 0.015), spelling (p = 0.009) and phonological decoding (0.017)- but not in the other big dataset analysed, CLDRC-RD. However, this dataset showed loadings generally concordant with those observed in UK-RD, with minor allele A having a positive effect on reading and language traits. Another significant association was detected with expressive language in SLIC (p = 0.002), but with an opposite allelic trend compared to the one seen in UK-RD, CLDRC-RD and in PC1 meta-analysis.

Our top association in *KIAA0319*, rs2143340 (Table S1g), showed generally comparable multivariate association loadings across all the traits analysed in all the datasets. Nominally

significant univariate associations were observed with phonological decoding (phonological choice p = 0.042), phoneme awareness (p = 0.024) and nonword repetition (p = 0.009) in CLDRC-RD; and with word reading (p = 0.008), spelling (0.003), phonological decoding (0.026) and orthographic coding (0.005) in UK-RD. Trends of association with word reading were observed also in CLDRC-RD and SLIC. Again, directions of effect were generally concordant with PC1/IQ-adjusted PC1 association, with minor allele G showing a negative effect on the traits.

Finally, rs2875891 (Table S1h), tagging associations in *ATP2C2*, was significantly associated with phoneme awareness in UK-RD (p = 0.04), and with word reading (p = 0.006), spelling (p = 0.018) and nonword repetition (p = 0.026) in SLIC. Similarly, we observed trends of association for word reading and spelling also in UK-RD. Minor allele T showed a positive effect on the traits in all the datasets except CLDRD-ADHD, in line with the allelic trend reported in PC1/IQ-adjusted PC1 meta-analysis.

Discussion

In the present chapter, we assessed associations for nine candidate genes implicated in RD and SLI by at least two previous studies: *MRPL19/GCFC2* (2p12), *ROBO1* (3p12), *DCDC2* (6p22), *KIAA0319* (6p22), *FOXP2* (7q31), *CNTNAP2* (7q35), *DYX1C1* (15q21), *CMIP* (16q23) and *ATP2C2* (16q24). Initially we focused on specific candidate SNPs in these genes, which had shown evidence of association with reading and language traits in previous studies. An assessment of 25 SNP associations supported by two or more independent studies revealed significant associations for SNPs within or close to *KIAA0319*, more specifically for rs2143340, rs3212236, rs9461045 and rs761100. All of these SNPs showed directions of effect consistent with those detected in the original studies (Francks et al.; 2004; Harold et al. 2006; Dennis et al., 2009). Among these SNPs, the strongest association was detected at rs2143340, located ~13 kb upstream of *KIAA0319*, within *TDP2* (tyrosyl-DNA phosphodiesterase 2, also known as *TTRAP*, TRAF and TNF receptor associated protein). The remaining associated SNPs in this region -rs3212236, rs9461045 and rs761100- were all located in intron 1, in the putative promoter of *KIAA0319*. All these SNPs were in relatively high LD ($r^2 > 0.63$), with the exception of rs761100 ($r^2 < 0.3$). This suggests the presence of two independent association signals in the putative promoter region of this gene: one tagged by rs761100, and the other one tagged by the haplotype rs2143340-rs3212236-rs9461045.

The latter association may be related to an altered expression of the gene, which finally leads to poor reading/language performance, as suggested by the association between rs9461045 and reduced *KIAA0319* expression (Dennis et al., 2009). The most associated SNP in this haplotype, rs2143340, revealed a broad pattern of effects on multiple reading/language traits, including both strictly reading-related measures (e.g. word reading and spelling) and skills more relevant to oral language (e.g. nonword repetition). This lends further support to the hypothesis that *KIAA0319* variants have pleiotropic effects on different reading and language skills, as already suggested by associations with several reading-related phenotypes (Franck et al., 2004; Paracchini et al., 2008); SLI status (Rice et al., 2009) and oral language ability (Newbury et al., 2011). This hypothesis is further corroborated by the cross-phenotypic effects of rs761100, which we detected in our analysis. For this SNP, however, significant genetic effects were mainly limited to word reading, spelling and phonological decoding in a single dataset (UK-RD).

In the set of 25 candidate SNPs initially assessed, some of the *ATP2C2* polymorphisms tested -namely rs16973771, rs2875891, rs8045507- showed nominally significant associations with PC1 and IQ-adjusted PC1 in our meta-analysis. Also in this case the allelic trends were generally concordant with those detected in the original report (Newbury et al., 2009). The most significantly associated SNP at this locus, rs2875891, revealed significant associations with word reading, spelling and nonword repetition in the SLIC dataset. This extends the range of genetic effects of *ATP2C2* variants -which had been detected so far only on nonword repetition and expressive/receptive language (Newbury et al., 2009)- to measures of reading and spelling, in contrast with the lack of association reported with these traits by previous studies (Newbury et al., 2011; Scerri et al., 2011).

An assessment of five additional SNPs in *ROBO1* and *FOXP2*, which had shown highly significant associations in the previous literature (Bates et al., 2011; Peter et al., 2011; Tran et al., 2014) but had not yet been replicated, revealed three significant associations within *ROBO1*, at rs6803202, rs4535189 and rs12495133.

rs6803202 and rs4535189 have been reported to be significantly associated with a phonological short term memory measure, i.e. nonword repetition, in a large population-based cohort from Australia (Bates et al., 2011). Nonetheless, their associations with PC1 and IQ-adjusted PC1 in our study showed opposite directions of effect compared to the original report (Bates et al., 2011). The assessment of cross-phenotypic associations of rs4535189 -tagging this association signal- revealed genetic effects on several reading and language

traits, although not always consistently across datasets. These elements suggest caution in the interpretation of this result.

rs12495133 has been recently associated with dyslexia in a family-based analysis of two RD Canadian datasets (Tran et al., 2014). In this study, the major (C) allele showed a significant overtransmission to RD cases, although the association survived correction for multiple testing only in one dataset. Conversely to rs6803202 and rs4535189, in our study rs12495133 was strongly associated ($p \sim 10^{-4}$) both with PC1 and IQ-adjusted PC1, and the allelic trend was consistent with the original report (Tran et al., 2014). This association would be significant even after an overly conservative correction for multiple testing of two traits (PC1 and IQ-adjusted PC1) and 30 candidate SNPs assessed in this study ($\alpha = 8.3 \times 10^{-4}$). As the original association was detected in a dataset other than those involved in our study, our result provides independent statistical support to this finding. The effect that we detected for rs12495133 went well beyond RD and strictly related traits, as suggested by the pleiotropic patterns of association of this SNP with several reading and language phenotypes. This was noticeable especially in our largest cohorts, CLDRC-RD and UK-RD.

rs12495133 is located within a putative enhancer region in *ROBO1* and is predicted to affect the binding of SOX5, a transcription factor important for the regulation of neuronal development and of axonal projections in cortical neurons (Kwan et al., 2008, Tran et al., 2014). This is consistent with the prominent axonal guidance role of ROBO1, in response to the chemo-attractant action of SLIT proteins (Seeger et al., 1993; Kidd et al., 1998; Andrews et al., 2006; 2008). rs12495133 is located more than 500 kb far from the top associations reported by Bates et al. (2011), rs6803202 and rs4535189, and is in low LD with them ($r^2 < 0.2$). Therefore, it is likely to represent an independent association signal.

When we extended the analysis to 13,827 SNPs which were lying within our candidate genes or in their putative regulatory regions, we detected no associations withstanding multiple testing correction, but found association signals supporting the results of the candidate SNP assessment. Indeed, the most significant association with PC1 was detected at rs3181234, which is located ~4 kb far from rs2143340 in *KIAA0319*, and is in high LD with it ($r^2 > 0.8$, see Figure 1a), likely tagging the same genetic effect. Similarly, the local top association with PC1 in *ROBO1* (rs1383407) is in high LD with rs12495133 ($r^2 > 0.9$, see Figure 2a), supporting the presence of a genetic effect in that region. On the other hand, the most significant association with IQ-adjusted PC1, rs2311350 in *ROBO1*, showed low LD with rs12495133 ($r^2 < 0.1$, see Figure 2b). This lends further support to the hypothesis of multiple

independent genetic effects on reading and language skills in *ROBO1* (see above). More importantly, the gene-wide analysis revealed an enrichment of SNP associations for *KIAA0319* and *ROBO1*, suggesting a pleiotropic effect of these genes on reading and language traits. Further support to the influence of *KIAA0319* and *ROBO1* on these traits came from the gene-based association analysis, which revealed significant associations only for these two genes.

Apart from the independent association that we report for rs12495133, our results for the majority of the 30 candidate SNPs assessed (presented in Table 1) can not be interpreted as being independent from previous findings. Most of these SNPs have been already tested in subsets of the datasets involved in our study, and sometimes the original associations were detected in these datasets. Nonetheless, our main aim here was not to produce independent replications of these findings, but rather to assess the cross-phenotypic effects of these candidate SNPs on several reading and language skills, in large datasets.

Surprisingly, our assessment provides no evidence of association with PC1/IQ-adjusted PC1 for the majority of candidate SNPs assessed in Table 1 and 2, and among these SNPs only associations of rs3212236 and rs9461045 with PC1 and of rs2143340 and rs12495133 with both PC traits would survive a hypothetical correction for multiple testing (see above). Even in the subsequent cross-phenotypic association analysis of the top associated SNPs, associations with individual traits were not always consistent across datasets. Two GWAS which were aimed at identifying genetic variants with pleiotropic effects on reading and language have already attempted to replicate these findings, but reported scarce evidence of replication for the candidate SNPs and genes tested here (Luciano et al., 2013; Eicher et al., 2013). Similarly, candidate SNP association analyses on large datasets -with sample size comprised between 500 and 2,000- have often led to weak or no replications of these associations, both in population-based cohorts (Luciano et al. 2007; Paracchini et al., 2011) and in clinical samples (Becker et al., 2014; Tran et al., 2014). This raises doubts on the replicability of the original findings, which were mainly generated through analysis of smaller samples. The lack of consistency of allelic effects across different studies reporting significant associations and the partially inconclusive results of subsequent meta-analyses (Zou et al., 2012; Tran et al., 2013; Zhong et al., 2013) further call into question most of the original associations assessed here.

These apparently contrasting results can be explained through several reasons. First, the heterogeneity of recruitment of samples analysed may lead to discrepant results across

different studies: some genetic variants may have stronger effects in the lower tail of the reading and language skills distributions (i.e. in RD and SLI selected samples) and negligible effects in a broader range of variation (i.e. in general population samples). Second, the heterogeneity of assessment of the phenotypes may result in traits that are supposed to tap into the same cognitive domain but actually represent slightly different abilities. This applies not only to continuous reading and language measures, but also to the classification of RD/SLI cases and controls, for which a consensus is far from being reached in the scientific community (Pennington & Bishop, 2009; Peterson & Pennington, 2012; Raskind et al., 2013). Third, different genetic backgrounds of the populations analysed may be a factor when comparing or meta-analyzing different association studies. The haplotype structure in a specific region may differ between populations, and so may change the LD between the tag SNP (where the association is detected) and the genuine causal SNP (which determines the association). In the presence of substantial population stratification this could even result in contrasting directions of effect for the same SNP in different studies (Lin et al., 2007; Luciano et al., 2007). Fourth, the irreproducibility of association studies may be due to type I errors, since false-positive results may easily occur in analyses of relatively small samples (Colhoun et al., 2003). While this may be a less likely explanation for those associated SNPs which have been functionally investigated -such as rs9461045 (as explained above)- it may reasonably account for spurious associations, which are more likely to be affected by publication biases (i.e. significant results tend to be favored for publication) and reporting biases (i.e. investigators tend to report only positive findings).

To try to solve these issues, different strategies may be adopted, such as i) increasing the homogeneity of association studies, ii) finding further support for statistical associations at the molecular level and iii) reducing the publication bias in the field. The former goal may be simply achieved by trying to use homogeneous inclusion and diagnostic criteria in the studies, as well as universal psychometric tests to assess the different cognitive traits. The second goal can be reached through the use of molecular biology techniques to functionally characterize variants identified in association studies, as successfully done for *ROBO1* (Hannula-Jouppi et al., 2005) and *KIAA0319* SNPs (Dennis et al., 2009). Such functional studies may help to elucidate the role of rs12495133 in *ROBO1*, for which we provide independent statistical support for association with reading and language traits. Last, reporting and organizing even negative findings and inconsistent associations into databases, as done by Bohland and colleagues (2014; http://neurospeech.org/sldb), will help to shed a

light on spurious associations and will constitute a valuable resource for future meta-analyses. These initiatives may help the scientific community to clarify the role of the candidate RD/SLI genes assessed here and, more in general, the genetic underpinnings of reading and language.

Acknowledgements

The authors declare no conflicts of interest. This work was supported by the Max Planck Society, the University of St Andrews, the EU (Neurodys, 018696), and the US National Institutes of Health (Grant ref: P50 HD027802). Genotyping at the Wellcome Trust Centre for Human Genetics was supported by the Wellcome Trust (090532/Z/09/Z) and a Medical Research Council Hub Grant (G0900747 91070). Silvia Paracchini is a Royal Society University Research Fellow. Dianne Newbury is an MRC Career Development Fellow and a Junior Research Fellow at St John's College, University of Oxford. Members of the SLI Consortium: *Wellcome Trust Centre for Human Genetics, Oxford:* D. F. Newbury, N. H. Simpson, R. Nudel, A. P. Monaco; *Max Planck Institute for Psycholinguistics, Nijmegen:* S. E. Fisher, C. Francks; *Newcomen Centre, Guy's Hospital, London:* G. Baird, V. Slonims, K Dworzynski; *Child and Adolescent Psychiatry Department and Medical Research Council Centre for Social, Developmental, and Genetic Psychiatry, Institute of Psychiatry, London:* P. F. Bolton; *Medical Research Council Centre for Social, Developmental, and Genetic Psychiatry Institute of Psychiatry, London:* E. Simonoff; *Department of Reproductive and Developmental Sciences, University of Edinburgh:* A. O'Hare; *Molecular Medicine Centre, University of Edinburgh:* J. Seckl; *Department of Speech and Language Therapy, Royal Hospital for Sick Children, Edinburgh*: H. Cowie; *Speech and Hearing Sciences, Queen Margaret University College:* A. Clark and J. Watson; *Department of Educational and Professional Studies, University of Strathclyde:* W. Cohen; *Department of Child Health, the University of Aberdeen*: A. Everitt, E. R. Hennessy, D. Shaw, P. J. Helms; *Audiology and Deafness, School of Psychological Sciences,University of Manchester:* Z. Simkin, G. Conti-Ramsden; *Department of Experimental Psychology, University of Oxford:* D. V. M. Bishop; *Biostatistics Department, Institute of Psychiatry, London:* A. Pickles.

Supplementary Material

- *S1: Supplementary Results.* Contribution of each dataset to the strength of the association of rs12495133 with PC1/IQ-adjusted PC1. Result of the gene-wide assessment of 13,827 SNPs in nine candidate RD/SLI genes. Cross-phenotypic effects of the candidate SNPs rs12495133, rs4535189 (*ROBO1*), rs761100, rs2143340 (*KIAA0319*) and rs2875891 (*ATP2C2*) on the different reading and language traits.

References

Adler, W.T., Platt, M.P., Mehlhorn, A.J., Haight, J.L., Currier, T.A., Etchegaray, M.A., Galaburda, A.M. & Rosen, G.D. (2013) Position of Neocortical Neurons Transfected at Different Gestational Ages with shRNA Targeted against Candidate Dyslexia Susceptibility Genes. *PLoS ONE,* **8,** e65179.

Alarcon, M., Abrahams, B., Stone, J., Duvall, J., Perederiy, J., Bomar, J., Sebat, J., Wigler, M., Martin, C., Ledbetter, D., Nelson, S., Cantor, R. & Geschwind, D. (2008) Linkage, association, and gene-expression analyses identify CNTNAP2 as an autism-susceptibility gene. *Am J Hum Genet,* **82,** 150 - 159.

Andrews, W., Barber, M., Hernadez-Miranda, L.R., Xian, J., Rakic, S., Sundaresan, V., Rabbitts, T.H., Pannell, R., Rabbitts, P., Thompson, H., Erskine, L., Murakami, F. & Parnavelas, J.G. (2008) The role of Slit-Robo signaling in the generation, migration and morphological differentiation of cortical interneurons. *Dev Biol,* **313,** 648-658.

Andrews, W., Liapi, A., Plachez, C., Camurri, L., Zhang, J., Mori, S., Murakami, F., Parnavelas, J.G., Sundaresan, V. & Richards, L.J. (2006) Robo1 regulates the development of major axon tracts and interneuron migration in the forebrain. *Development,* **133,** 2243-2252.

Anney, R. & Klei, L. & Pinto, D. & Almeida, J. & Bacchelli, E. & Baird, G. & Bolshakova, N. & Bölte, S. & Bolton, P.F. & Bourgeron, T. & Brennan, S. & Brian, J. & Casey, J. & Conroy, J. & Correia, C. & Corsello, C. & Crawford, E.L. & de Jonge, M. & Delorme, R. & Duketis, E. & Duque, F. & Estes, A. & Farrar, P. & Fernandez, B.A. & Folstein, S.E. & Fombonne, E. & Gilbert, J. & Gillberg, C. & Glessner, J.T. & Green, A. & Green, J. & Guter, S.J. & Heron, E.A. & Holt, R. & Howe, J.L. & Hughes, G. & Hus, V. & Igliozzi, R. & Jacob, S. & Kenny, G.P. & Kim, C. & Kolevzon, A. & Kustanovich, V. & Lajonchere, C.M. & Lamb, J.A. & Law-Smith, M. & Leboyer, M. & Le Couteur, A. & Leventhal, B.L. & Liu, X.-Q. & Lombard, F. & Lord, C. & Lotspeich, L. & Lund, S.C. & Magalhaes, T.R. & Mantoulan, C. & McDougle, C.J. & Melhem, N.M. & Merikangas, A. & Minshew, N.J. & Mirza, G.K. & Munson, J. & Noakes, C. & Nygren, G. & Papanikolaou, K. & Pagnamenta, A.T. & Parrini, B. & Paton, T. & Pickles, A. & Posey, D.J. & Poustka, F. & Ragoussis, J. & Regan, R. & Roberts, W. & Roeder, K. & Roge, B. & Rutter, M.L. & Schlitt, S. & Shah, N. & Sheffield, V.C. & Soorya, L. & Sousa, I. & Stoppioni, V. & Sykes, N. & Tancredi, R. & Thompson, A.P. & Thomson, S. & Tryfon, A. & Tsiantis, J. & Van Engeland, H. & Vincent, J.B. & Volkmar, F. & Vorstman, J. & Wallace, S. & Wing, K. & Wittemeyer, K. & Wood, S. & Zurawiecki, D. & Zwaigenbaum, L. & Bailey, A.J., et al. (2012) Individual common variants exert weak effects on the risk for autism spectrum disorders. *Hum Mol Genet,* **21,** 4781-4792.

Anthoni, H., Zucchelli, M., Matsson, H., Muller-Myhsok, B., Fransson, I., Schumacher, J., Massinen, S., Onkamo, P., Warnke, A., Griesemann, H., Hoffmann, P., Nopola-Hemmi, J., Lyytinen, H., Schulte-Korne, G., Kere, J., Nothen, M.M. & Peyrard-Janvid, M. (2007) A locus on 2p12 containing the co-regulated MRPL19 and C2ORF3 genes is associated to dyslexia. *Hum Mol Genet,* **16,** 667-677.

Baek, Seung T., Kerjan, G., Bielas, Stephanie L., Lee, Ji E., Fenstermaker, Ali G., Novarino, G. & Gleeson, Joseph G. (2014) Off-Target Effect of doublecortin Family shRNA on Neuronal Migration Associated with Endogenous MicroRNA Dysregulation. *Neuron*, **82**, 1255-1262.

Bates, T.C., Luciano, M., Medland, S.E., Montgomery, G.W., Wright, M.J. & Martin, N.G. (2011) Genetic variance in a component of the language acquisition device: ROBO1 polymorphisms associated with phonological buffer deficits. *Behav Genet,* **41,** 50-57.

Becker, J., Czamara, D., Scerri, T.S., Ramus, F., Csepe, V., Talcott, J.B., Stein, J., Morris, A., Ludwig, K.U., Hoffmann, P., Honbolygo, F., Toth, D., Fauchereau, F., Bogliotti, C., Iannuzzi, S., Chaix, Y., Valdois, S., Billard, C., George, F., Soares-Boucaud, I., Gerard, C.L., van der Mark, S., Schulz, E., Vaessen, A., Maurer, U., Lohvansuu, K., Lyytinen, H., Zucchelli, M., Brandeis, D., Blomert, L., Leppanen, P.H., Bruder, J., Monaco, A.P., Muller-Myhsok, B., Kere, J., Landerl, K., Nothen, M.M., Schulte-Korne, G., Paracchini, S., Peyrard-Janvid, M. & Schumacher, J. (2014) Genetic analysis of dyslexia candidate genes in the European cross-linguistic NeuroDys cohort. *Eur J Hum Genet*, **22**, 675–680.

Bishop, D.V. (1994) Is specific language impairment a valid diagnostic category? Genetic and psycholinguistic evidence. *Philos Trans R Soc Lond B Biol Sci,* **346,** 105-111.

Bishop, D.V.M. & Hayiou-Thomas, M.E. (2008) Heritability of specific language impairment depends on

diagnostic criteria. *Genes Brain Behav,* **7,** 365-372.

Bishop, D.V.M. & Snowling, M.J. (2004) Developmental Dyslexia and Specific Language Impairment: Same or Different? *Psychol Bull,* **130,** 858-886.

Bohland, J.W., Myers, E.M. & Kim, E. (2014) An Informatics Approach to Integrating Genetic and Neurological Data in Speech and Language Neuroscience. *Neuroinformatics,* **12,** 39-62.

Carrion-Castillo, A., Franke, B. & Fisher, S.E. (2013) Molecular Genetics of Dyslexia: An Overview. *Dyslexia,* **19,** 214-240.

Chandrasekar, G., Vesterlund, L., Hultenby, K., Tapia-Páez, I. & Kere, J. (2013) The Zebrafish Orthologue of the Dyslexia Candidate Gene DYX1C1 Is Essential for Cilia Growth and Function. *PLoS ONE*, **8**, e63123.

Colhoun, H.M., McKeigue, P.M. & Smith, G.D. (2003) Problems of reporting genetic associations with complex outcomes. *Lancet,* **361,** 865-872.

Cope, N., Harold, D., Hill, G., Moskvina, V., Stevenson, J., Holmans, P., Owen, M.J., O'Donovan, M.C. & Williams, J. (2005) Strong Evidence That KIAA0319 on Chromosome 6p Is a Susceptibility Gene for Developmental Dyslexia. *Am J Hum Genet,* **76,** 581-591.

Dennis, M.Y., Paracchini, S., Scerri, T.S., Prokunina-Olsson, L., Knight, J.C., Wade-Martins, R., Coggill, P., Beck, S., Green, E.D. & Monaco, A.P. (2009) A Common Variant Associated with Dyslexia Reduces Expression of the KIAA0319 Gene. *PLoS Genet,* **5,** e1000436.

Eicher, J., Powers, N., Miller, L., Mueller, K., Mascheretti, S., Marino, C., Willcutt, E., DeFries, J., Olson, R., Smith, S., Pennington, B., Tomblin, J.B., Ring, S. & Gruen, J. (2014) Characterization of the DYX2 locus on chromosome 6p22 with reading disability, language impairment, and IQ. *Hum Genet,* **133,** 869-881.

Eicher, J.D., Powers, N.R., Miller, L.L., Akshoomoff, N., Amaral, D.G., Bloss, C.S., Libiger, O., Schork, N.J., Darst, B.F., Casey, B.J., Chang, L., Ernst, T., Frazier, J., Kaufmann, W.E., Keating, B., Kenet, T., Kennedy, D., Mostofsky, S., Murray, S.S., Sowell, E.R., Bartsch, H., Kuperman, J.M., Brown, T.T., Hagler, D.J., Jr., Dale, A.M., Jernigan, T.L., Pourcain, B.S., Smith, G.D., Ring, S.M. & Gruen, J.R. (2013) Genome-Wide Association Study of Shared Components of Reading Disability and Language Impairment. *Genes Brain Behav*, **8,** 792-801.

Ferreira, M.A. & Purcell, S.M. (2009) A multivariate test of association. *Bioinformatics,* **25,** 132-133.

Fisher, S.E. & DeFries, J.C. (2002) Developmental dyslexia: genetic dissection of a complex cognitive trait. *Nat Rev Neurosci,* **3,** 767-780.

Fisher, S.E. & Scharff, C. (2009) FOXP2 as a molecular window into speech and language. *Trends Genet,* **25,** 166-177.

Francks, C., Paracchini, S., Smith, S.D., Richardson, A.J., Scerri, T.S., Cardon, L.R., Marlow, A.J., MacPhie, I.L., Walter, J., Pennington, B.F., Fisher, S.E., Olson, R.K., DeFries, J.C., Stein, J.F. & Monaco, A.P. (2004) A 77-Kilobase Region of Chromosome 6p22.2 Is Associated with Dyslexia in Families From the United Kingdom and From the United States. *Am J Hum Genet,* **75,** 1046-1058.

Galaburda, A.M. & Cestnick, L. (2003) Developmental dyslexia. *Rev Neurol,* **36,** 3-9.

Good, C.D., Johnsrude, I., Ashburner, J., Henson, R.N., Friston, K.J. & Frackowiak, R.S. (2001) Cerebral asymmetry and the effects of sex and handedness on brain structure: a voxel-based morphometric analysis of 465 normal adult human brains. *Neuroimage,* **14,** 685-700.

Graham, S.A. & Fisher, S.E. (2013) Decoding the genetics of speech and language. *Curr Opin Neurobiol,* **23,** 43-51.

Grati, M.h., Chakchouk, I., Ma, Q., Bensaid, M., Desmidt, A., Turki, N., Yan, D., Baanannou, A., Mittal, R., Driss, N., Blanton, S., Farooq, A., Lu, Z., Liu, X.Z. & Masmoudi, S. (2015) A missense mutation in DCDC2 causes human recessive deafness DFNB66, likely by interfering with sensory hair cell and supporting cell cilia length regulation. *Hum Molec Genet*, 10.1093/hmg/ddv009.

Grimbert, P., Valanciute, A., Audard, V., Lang, P., Guellaen, G. & Sahali, D. (2004) The Filamin-A is a partner of Tc-mip, a new adapter protein involved in c-maf-dependent Th2 signaling pathway. *Mol Immunol,* **40,** 1257-1261.

Hannula-Jouppi, K., Kaminen-Ahola, N., Taipale, M., Eklund, R., Nopola-Hemmi, J., Kääriäinen, H. & Kere, J. (2005) The Axon Guidance Receptor Gene *ROBO1* Is a Candidate Gene for Developmental Dyslexia.

PLoS Genet, **1,** e50.

Harold, D., Paracchini, S., Scerri, T., Dennis, M., Cope, N., Hill, G., Moskvina, V., Walter, J., Richardson, A.J., Owen, M.J., Stein, J.F., Green, E.D., O'Donovan, M.C., Williams, J. & Monaco, A.P. (2006) Further evidence that the KIAA0319 gene confers susceptibility to developmental dyslexia. *Mol Psychiatry,* **11,** 1085-1091.

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet,* **44,** 955-959.

Kidd, T., Brose, K., Mitchell, K.J., Fetter, R.D., Tessier-Lavigne, M., Goodman, C.S. & Tear, G. (1998) Roundabout Controls Axon Crossing of the CNS Midline and Defines a Novel Subfamily of Evolutionarily Conserved Guidance Receptors. *Cell,* **92,** 205-215.

Kwan, K.Y., Lam, M.M.S., Krsnik, Ž., Kawasawa, Y.I., Lefebvre, V. & Šestan, N. (2008) SOX5 postmitotically regulates migration, postmigratory differentiation, and projections of subplate and deep-layer neocortical neurons. *Proc Natl Acad Sci USA,* **105,** 16021-16026.

Lai, C.S.L., Fisher, S.E., Hurst, J.A., Vargha-Khadem, F. & Monaco, A.P. (2001) A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature,* **413,** 519-523.

Li, M.X., Yeung, J.Y., Cherny, S. & Sham, P. (2012) Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum Genet*, **131**, 747-756.

Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol,* **34,** 816-834.

Lin, P.I., Vance, J.M., Pericak-Vance, M.A. & Martin, E.R. (2007) No Gene Is an Island: The Flip-Flop Phenomenon. *Am J Hum Genet,* **80,** 531-538.

Liu, J.Z., McRae, A.F., Nyholt, D.R., Medland, S.E., Wray, N.R., Brown, K.M., Hayward, N.K., Montgomery, G.W., Visscher, P.M., Martin, N.G. & Macgregor, S. (2010) A versatile gene-based test for genome-wide association studies. *Am J Hum Genet,* **87,** 139-145.

Lombardo, M.V., Ashwin, E., Auyeung, B., Chakrabarti, B., Taylor, K., Hackett, G., Bullmore, E.T. & Baron-Cohen, S. (2012) Fetal testosterone influences sexually dimorphic gray matter in the human brain. *J Neurosci,* **32,** 674-680.

Luciano, M., Evans, D.M., Hansell, N.K., Medland, S.E., Montgomery, G.W., Martin, N.G., Wright, M.J. & Bates, T.C. (2013) A genome-wide association study for reading and language abilities in two population cohorts. *Genes Brain Behav,* **12,** 645-652.

Luciano, M., Lind, P.A., Duffy, D.L., Castles, A., Wright, M.J., Montgomery, G.W., Martin, N.G. & Bates, T.C. (2007) A haplotype spanning KIAA0319 and TTRAP is associated with normal variation in reading and spelling ability. *Biol Psychiatry,* **62,** 811-817.

MacDermot, K.D., Bonora, E., Sykes, N., Coupe, A.M., Lai, C.S., Vernes, S.C., Vargha-Khadem, F., McKenzie, F., Smith, R.L., Monaco, A.P. & Fisher, S.E. (2005) Identification of FOXP2 truncation as a novel cause of developmental speech and language deficits. *Am J Hum Genet,* **76,** 1074-1080.

Marino, C., Meng, H., Mascheretti, S., Rusconi, M., Cope, N., Giorda, R., Molteni, M. & Gruen, J.R. (2012) DCDC2 genetic variants and susceptibility to developmental dyslexia. *Psychiatr Genet,* **22,** 25-30.

Massinen, S., Hokkanen, M.-E., Matsson, H., Tammimies, K., Tapia-Páez, I., Dahlström-Heuser, V., Kuja-Panula, J., Burghoorn, J., Jeppsson, K.E., Swoboda, P., Peyrard-Janvid, M., Toftgård, R., Castrén, E. & Kere, J. (2011) Increased Expression of the Dyslexia Candidate Gene DCDC2 Affects Length and Signaling of Primary Cilia in Neurons. *PLoS ONE,* **6,** e20580.

McArthur, G.M., Hogben, J.H., Edwards, V.T., Heath, S.M. & Mengler, E.D. (2000) On the "specifics" of specific reading disability and specific language impairment. *J Child Psychol Psychiatry,* **41,** 869–874.

Meng, H., Powers, N.R., Tang, L., Cope, N.A., Zhang, P.X., Fuleihan, R., Gibson, C., Page, G.P. & Gruen, J.R. (2011) A dyslexia-associated variant in DCDC2 changes gene expression. *Behav Genet,* **41,** 58-66.

Meng, H., Smith, S.D., Hager, K., Held, M., Liu, J., Olson, R.K., Pennington, B.F., DeFries, J.C., Gelernter, J., O'Reilly-Pol, T., Somlo, S., Skudlarski, P., Shaywitz, S.E., Shaywitz, B.A., Marchione, K., Wang, Y., Paramasivam, M., LoTurco, J.J., Page, G.P. & Gruen, J.R. (2005) DCDC2 is associated with reading disability and modulates neuronal development in the brain. *Proc Natl Acad Sci USA,* **102,** 17053-17058.

Newbury, D., Fisher, S. & Monaco, A. (2010) Recent advances in the genetics of language impairment. *Genome Med,* **2,** 6.

Newbury, D.F. & Monaco, A.P. (2010) Genetic advances in the study of speech and language disorders. *Neuron,* **68,** 309-320.

Newbury, D.F., Paracchini, S., Scerri, T.S., Winchester, L., Addis, L., Richardson, A.J., Walter, J., Stein, J.F., Talcott, J.B. & Monaco, A.P. (2011) Investigation of dyslexia and SLI risk variants in reading- and language-impaired subjects. *Behav Genet,* **41,** 90-104.

Newbury, D.F., Winchester, L., Addis, L., Paracchini, S., Buckingham, L.L., Clark, A., Cohen, W., Cowie, H., Dworzynski, K., Everitt, A., Goodyer, I.M., Hennessy, E., Kindley, A.D., Miller, L.L., Nasir, J., O'Hare, A., Shaw, D., Simkin, Z., Simonoff, E., Slonims, V., Watson, J., Ragoussis, J., Fisher, S.E., Seckl, J.R., Helms, P.J., Bolton, P.F., Pickles, A., Conti-Ramsden, G., Baird, G., Bishop, D.V. & Monaco, A.P. (2009) CMIP and ATP2C2 modulate phonological short-term memory in language impairment. *Am J Hum Genet,* **85,** 264-272.

Nopola-Hemmi, J., Myllyluoma, B., Haltia, T., Taipale, M., Ollikainen, V., Ahonen, T., Voutilainen, A., Kere, J. & Widén, E. (2001) A dominant gene for developmental dyslexia on chromosome 3. *J Med Genet,* **38,** 658-664.

Nopola-Hemmi, J., Taipale, M., Haltia, T., Lehesjoki, A.-E., Voutilainen, A. & Kere, J. (2000) Two translocations of chromosome 15q associated with dyslexia. *J Med Genet,* **37,** 771-775.

Paracchini, S., Ang, Q.W., Stanley, F.J., Monaco, A.P., Pennell, C.E. & Whitehouse, A.J. (2011) Analysis of dyslexia candidate genes in the Raine cohort representing the general Australian population. *Genes Brain Behav,* **10,** 158-165.

Paracchini, S., Steer, C., Buckingham, L.L., Morris, A., Ring, S., Scerri, T., Stein, J., Pembrey, M., Ragoussis, J., Golding, J. & Monaco, A. (2008) Association of the KIAA0319 Dyslexia Susceptibility Gene With Reading Skills in the General Population. *Am J Psychiatry,* **165,** 1576-1584.

Pennington, B.F. & Bishop, D.V. (2009) Relations among speech, language, and reading disorders. *Annu Rev Psych,* **60,** 283-306.

Peschansky, V.J., Burbridge, T.J., Volz, A.J., Fiondella, C., Wissner-Gross, Z., Galaburda, A.M., Turco, J.J.L. & Rosen, G.D. (2010) The Effect of Variation in Expression of the Candidate Dyslexia Susceptibility Gene Homolog Kiaa0319 on Neuronal Migration and Dendritic Morphology in the Rat. *Cereb Cortex,* **20,** 884-897.

Peter, B., Raskind, W., Matsushita, M., Lisowski, M., Vu, T., Berninger, V., Wijsman, E. & Brkanac, Z. (2011) Replication of CNTNAP2 association with nonword repetition and support for FOXP2 association with timed reading and motor activities in a dyslexia family sample. *J Neurodevelop Disord,* **3,** 39-49.

Peterson, R.L. & Pennington, B.F. (2012) Developmental dyslexia. *Lancet,* **379,** 1997-2007.

Powers, N.R., Eicher, J.D., Butter, F., Kong, Y., Miller, L.L., Ring, S.M., Mann, M. & Gruen, J.R. (2013) Alleles of a polymorphic ETV6 binding site in DCDC2 confer risk of reading and language impairment. *Am J Hum Genet,* **93,** 19-28.

Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R. & Willer, C.J. (2010) LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics,* **26,** 2336-2337.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. & Sham, P.C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet,* **81,** 559-575.

Raskind, W.H., Peter, B., Richards, T., Eckert, M.M. & Berninger, V.W. (2013) The genetics of reading disabilities: from phenotypes to candidate genes. *Front Psychol,* **3,** 601.

Rice, M., Smith, S. & Gayán, J. (2009) Convergent genetic linkage and associations to language, speech and reading measures in families of probands with Specific Language Impairment. *J Neurodevelop Disord,* **1,** 264-282.

Rodenas-Cuadrado, P., Ho, J. & Vernes, S.C. (2014) Shining a light on CNTNAP2: complex functions to complex disorders. *Eur J Hum Genet,* **22,** 171-178.

Scerri, T.S., Morris, A.P., Buckingham, L.L., Newbury, D.F., Miller, L.L., Monaco, A.P., Bishop, D.V.M. & Paracchini, S. (2011) DCDC2, KIAA0319 and CMIP Are Associated with Reading-Related Traits. *Biol*

Psychiatry, **70,** 237-245.

Schueler, M., Braun, Daniela A., Chandrasekar, G., Gee, Heon Y., Klasson, Timothy D., Halbritter, J., Bieder, A., Porath, Jonathan D., Airik, R., Zhou, W., LoTurco, Joseph J., Che, A., Otto, Edgar A., Böckenhauer, D., Sebire, Neil J., Honzik, T., Harris, Peter C., Koon, Sarah J., Gunay-Aygun, M., Saunier, S., Zerres, K., Bruechle, Nadina O., Drenth, Joost P.H., Pelletier, L., Tapia-Páez, I., Lifton, Richard P., Giles, Rachel H., Kere, J. & Hildebrandt, F. (2015) DCDC2 Mutations Cause a Renal-Hepatic Ciliopathy by Disrupting Wnt Signaling. *Am J Hum Genet*, **96**, 81-92.

Schumacher, J., Anthoni, H., Dahdouh, F., König, I.R., Hillmer, A.M., Kluck, N., Manthey, M., Plume, E., Warnke, A., Remschmidt, H., Hülsmann, J., Cichon, S., Lindgren, C.M., Propping, P., Zucchelli, M., Ziegler, A., Peyrard-Janvid, M., Schulte-Körne, G., Nöthen, M.M. & Kere, J. (2006) Strong Genetic Evidence of DCDC2 as a Susceptibility Gene for Dyslexia. *Am J Hum Genet,* **78,** 52-62.

Seeger, M., Tear, G., Ferres-Marco, D. & Goodman, C.S. (1993) Mutations affecting growth cone guidance in drosophila: Genes necessary for guidance toward or away from the midline. *Neuron,* **10,** 409-426.

Shapleske, J., Rossell, S.L., Woodruff, P.W.R. & David, A.S. (1999) The planum temporale: a systematic, quantitative review of its structural, functional and clinical significance. *Brain Res Rev,* **29,** 26-49.

Shaywitz, S.E., Shaywitz, B.A., Fletcher, J.M.à. & Escobar, M.D. (1990) Prevalence of reading disability in boys and girls. Results of the Connecticut Longitudinal Study. *JAMA,* **264,** 998–1002.

Snowling, M., Bishop, D.V. & Stothard, S.E. (2000) Is preschool language impairment a risk factor for dyslexia in adolescence? *J Child Psychol Psychiatry. ,* **41,** 587–600.

Taipale, M., Kaminen, N., Nopola-Hemmi, J., Haltia, T., Myllyluoma, B., Lyytinen, H., Muller, K., Kaaranen, M., Lindsberg, P.J., Hannula-Jouppi, K. & Kere, J. (2003) A candidate gene for developmental dyslexia encodes a nuclear tetratricopeptide repeat domain protein dynamically regulated in brain. *Proc Natl Acad Sci USA,* **100,** 11553-11558.

Tammimies, K., Vitezic, M., Matsson, H., Le Guyader, S., Burglin, T.R., Ohman, T., Stromblad, S., Daub, C.O., Nyman, T.A., Kere, J. & Tapia-Paez, I. (2013) Molecular networks of DYX1C1 gene show connection to neuronal migration genes and cytoskeletal proteins. *Biol Psychiatry,* **73,** 583-590.

Tarkar, A., Loges, N.T., Slagle, C.E., Francis, R., Dougherty, G.W., Tamayo, J.V., Shook, B., Cantino, M., Schwartz, D., Jahnke, C., Olbrich, H., Werner, C., Raidt, J., Pennekamp, P., Abouhamed, M., Hjeij, R., Kohler, G., Griese, M., Li, Y., Lemke, K., Klena, N., Liu, X., Gabriel, G., Tobita, K., Jaspers, M., Morgan, L.C., Shapiro, A.J., Letteboer, S.J.F., Mans, D.A., Carson, J.L., Leigh, M.W., Wolf, W.E., Chen, S., Lucas, J.S., Onoufriadis, A., Plagnol, V., Schmidts, M., Boldt, K., Uk10K, Roepman, R., Zariwala, M.A., Lo, C.W., Mitchison, H.M., Knowles, M.R., Burdine, R.D., LoTurco, J.J. & Omran, H. (2013) DYX1C1 is required for axonemal dynein assembly and ciliary motility. *Nat Genet*, **45**, 995-1003.

The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature,* **491,** 56-65.

The International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. *Nature,* **467,** 52-58.

Toma, C., Hervas, A., Torrico, B., Balmana, N., Salgado, M., Maristany, M., Vilella, E., Martinez-Leal, R., Planelles, M.I., Cusco, I., del Campo, M., Perez-Jurado, L.A., Caballero-Andaluz, R., de Diego-Otero, Y., Perez-Costillas, L., Ramos-Quiroga, J.A., Ribases, M., Bayes, M. & Cormand, B. (2013) Analysis of two language-related genes in autism: a case-control association study of FOXP2 and CNTNAP2. *Psychiatr Genet*, **23**, 82-85.

Tomblin, J.B., Records, N.L., Buckwalter, P., Zhang, X., Smith, E. & O'Brien, M. (1997) Prevalence of specific language impairment in kindergarten children. *J Speech Lang Hear Res,* **40,** 1245-1260.

Tran, C., Gagnon, F., Wigg, K.G., Feng, Y., Gomez, L., Cate-Carter, T.D., Kerr, E.N., Field, L.L., Kaplan, B.J., Lovett, M.W. & Barr, C.L. (2013) A family-based association analysis and meta-analysis of the reading disabilities candidate gene DYX1C1. *Am J Med Genet B Neuropsychiatr Genet,* **162,** 146-156.

Tran, C., Wigg, K.G., Zhang, K., Cate-Carter, T.D., Kerr, E., Field, L.L., Kaplan, B.J., Lovett, M.W. & Barr, C.L. (2014) Association of the ROBO1 gene with reading disabilities in a family-based analysis. *Genes Brain Behav,* **13,** 430-438.

Velayos-Baeza, A., Toma, C., da Roza, S., Paracchini, S. & Monaco, A. (2007) Alternative splicing in the

dyslexia-associated gene KIAA0319. *Mamm Genome,* **18,** 627-634.

Velayos-Baeza, A., Toma, C., Paracchini, S. & Monaco, A.P. (2008) The dyslexia-associated gene KIAA0319 encodes highly N- and O-glycosylated plasma membrane and secreted isoforms. *Hum Mol Genet,* **17,** 859-871.

Venkatesh, S.K., Siddaiah, A., Padakannaya, P. & Ramachandra, N.B. (2013) Analysis of genetic variants of dyslexia candidate genes KIAA0319 and DCDC2 in Indian population. *J Hum Genet,* **58,** 531-538.

Vernes, S.C., Newbury, D.F., Abrahams, B.S., Winchester, L., Nicod, J., Groszer, M., Alarcon, M., Oliver, P.L., Davies, K.E., Geschwind, D.H., Monaco, A.P. & Fisher, S.E. (2008) A functional genetic link between distinct developmental language disorders. *N Engl J Med,* **359,** 2337-2345.

Vernes, S.C., Oliver, P.L., Spiteri, E., Lockstone, H.E., Puliyadi, R., Taylor, J.M., Ho, J., Mombereau, C., Brewer, A., Lowy, E., Nicod, J., Groszer, M., Baban, D., Sahgal, N., Cazier, J.-B., Ragoussis, J., Davies, K.E., Geschwind, D.H. & Fisher, S.E. (2011) Foxp2 Regulates Gene Networks Implicated in Neurite Outgrowth in the Developing Brain. *PLoS Genet,* **7,** e1002145.

Wang, Y., Yin, X., Rosen, G., Gabel, L., Guadiana, S.M., Sarkisian, M.R., Galaburda, A.M. & LoTurco, J.J. (2011) Dcdc2 knockout mice display exacerbated developmental disruptions following knockdown of doublecortin. *Neuroscience,* **190,** 398-408.

Whitehouse, A.J., Mattes, E., Maybery, M.T., Sawyer, M.G., Jacoby, P., Keelan, J.A. & Hickey, M. (2012) Sex-specific associations between umbilical cord blood testosterone levels and language delay in early childhood. *J Child Psychol Psychiatry,* **53,** 726-734.

Wilcke, A., Ligges, C., Burkhardt, J., Alexander, M., Wolf, C., Quente, E., Ahnert, P., Hoffmann, P., Becker, A., Muller-Myhsok, B., Cichon, S., Boltze, J. & Kirsten, H. (2012) Imaging genetics of FOXP2 in dyslexia. *Eur J Hum Genet,* **20,** 224-229.

Wilcke, A., Weissfuss, J., Kirsten, H., Wolfram, G., Boltze, J. & Ahnert, P. (2009) The role of gene DCDC2 in German dyslexics. *Ann Dyslexia,* **59,** 1-11.

Willer, C.J., Li, Y. & Abecasis, G.R. (2010) METAL: fast and efficient meta-analysis of genome wide association scans. *Bioinformatics,* **26,** 2190-2191.

Zhong, R., Yang, B., Tang, H., Zou, L., Song, R., Zhu, L.-Q. & Miao, X. (2013) Meta-analysis of the Association Between DCDC2 Polymorphisms and Risk of Dyslexia. *Mol Neurobiol,* **47,** 435-442.

Zou, L., Chen, W., Shao, S., Sun, Z., Zhong, R., Shi, J., Miao, X. & Song, R. (2012) Genetic variant in KIAA0319, but not in DYX1C1, is associated with risk of dyslexia: An integrated meta-analysis. *Am J Med Genet B Neuropsychiatr Genet,* **159B,** 970-976.

S1: Supplementary Results

S1a)

Trait	PC1			IQ-adjusted PC1		
Dataset	P-value	Weighted Z score	Beta[a]	P-value	Weighted Z score	Beta[a]
CLDRC-RD	0.049	1.06	0.126	0.037	1.14	0.134
UK-RD	0.023	1.6	0.122	0.016	1.67	0.126
SLIC	0.268	0.4	0.122	0.228	0.44	0.123
CLDRC-ADHD	0.314	0.3	0.13	0.347	0.28	0.123
Meta-Analysis	7.9×10^{-4}	3.36	NA[b]	4.1×10^{-4}	3.53	NA[b]

Table S1a. Contribution of each dataset to the strength of the association of rs12495133 with PC1 and IQ-adjusted PC1. This is represented by PLINK univariate QFAM p-value and beta regression coefficient for each association test (i.e. for each dataset), and by corresponding weighted Z-score, as computed by METAL sample size based algorithm (Willer et al. 2010). The sign of z scores and beta values refer to the effect of the minor allele (A).

[a] Although beta values computed by QFAM are not adjusted for family structure, they are reported in the table as a term of comparison of effect sizes across datasets. [b] Not Applicable, since the METAL sample size-based algorithm computes a global weighted Z score (but not a Beta coefficient).

S1b)

Chr	SNP[a]	Position	Allele1	Allele2	Freq (%)[b]	P-value	Direction[c]	HetPVal[d]	Gene
6	rs3181234	24655165	t	c	14.8	5.15×10^{-5}	----	0.71	KIAA0319
6	rs11961837	24681787	t	c	15.18	8.07×10^{-5}	----	0.74	KIAA0319
6	rs73392549	24677968	a	g	14.58	8.5×10^{-5}	----	0.7	KIAA0319
6	rs77272080	24679178	a	c	14.62	8.83×10^{-5}	----	0.68	KIAA0319
6	rs11962639	24683150	a	g	15.02	9.59×10^{-5}	----	0.66	KIAA0319
6	rs7768291	24656571	a	g	85.03	1.13×10^{-4}	++++	0.73	KIAA0319
6	rs6909884	24676985	a	g	14.58	1.16×10^{-4}	----	0.67	KIAA0319
6	rs2143340	24659071	a	g	85.34	1.44×10^{-4}	++++	0.74	KIAA0319
3	rs1383407	78908171	t	c	59.95	1.85×10^{-4}	----	0.93	ROBO1
6	rs3756819	24665340	a	c	85.39	2.14×10^{-4}	++++	0.79	KIAA0319
6	6:24657853:D	24657853	d	r	14.05	2.47×10^{-4}	----	0.73	KIAA0319
6	rs3777665	24693523	t	c	14.03	3.01×10^{-4}	----	0.6	KIAA0319
6	rs3181244	24651388	a	g	14.29	3.56×10^{-4}	----	0.69	KIAA0319
3	rs1159912	79194332	a	g	66.62	4.08×10^{-4}	+++-	0.41	ROBO1
6	rs146260219	24403411	t	c	96.93	4.17×10^{-4}	++++	0.08	DCDC2
6	rs114966185	24405067	t	c	96.91	4.55×10^{-4}	++++	0.1	DCDC2
3	rs1159913	79193997	a	g	66.62	5.03×10^{-4}	+++-	0.44	ROBO1
6	rs2876680	24679994	t	c	79.55	5.03×10^{-4}	++++	0.88	KIAA0319
6	rs146830531	24499285	t	c	97.49	5.35×10^{-4}	++++	0.95	KIAA0319
6	rs1923189	24682906	a	g	79.6	5.63×10^{-4}	++++	0.92	KIAA0319
3	rs60200150	79191634	a	c	66.74	5.71×10^{-4}	+++-	0.47	ROBO1
3	rs73114798	78927132	t	c	36.34	5.8×10^{-4}	++++	1	ROBO1
3	rs2311351	79197773	t	c	66.61	5.81×10^{-4}	+++-	0.45	ROBO1
3	rs2168373	79193739	c	g	66.62	6.39×10^{-4}	+++-	0.47	ROBO1
6	rs9467254	24683119	c	g	79.6	6.41×10^{-4}	++++	0.95	KIAA0319
3	rs11127636	78922852	a	c	59.34	6.66×10^{-4}	----	0.96	ROBO1
3	rs2311350	79197923	a	g	66.61	6.7×10^{-4}	+++-	0.47	ROBO1
3	rs12486635	79197710	a	g	66.61	6.72×10^{-4}	+++-	0.45	ROBO1
6	rs114221483	24497299	t	c	97.49	6.76×10^{-4}	++++	0.93	KIAA0319
3	rs2311349	79197994	t	c	66.61	6.98×10^{-4}	+++-	0.46	ROBO1
3	rs12495133	78921520	a	c	36.34	7.91×10^{-4}	++++	1	ROBO1
3	rs7638301	79196370	c	g	33.28	8.21×10^{-4}	---+	0.47	ROBO1
3	rs7631406	79190590	t	c	33	8.29×10^{-4}	---+	0.56	ROBO1
3	rs1378632	79195917	t	c	66.71	8.33×10^{-4}	+++-	0.44	ROBO1
3	rs4680943	78926431	a	t	59.44	8.64×10^{-4}	----	0.98	ROBO1
6	6:24641871:I	24641871	i	r	60.41	9.23×10^{-4}	--+-	0.2	KIAA0319
3	rs9823929	79184872	t	c	29.36	9.59×10^{-4}	---+	0.38	ROBO1
3	rs6548614	79178697	a	g	70.51	9.9×10^{-4}	+++-	0.42	ROBO1

S1c)

Chr	SNP[a]	Position	Allele1	Allele2	Freq (%)[b]	P-value	Direction[c]	HetPVal[d]	Gene
3	rs2311350	79197923	a	g	66.61	2.86×10^{-5}	+++-	0.38	ROBO1
3	rs2311349	79197994	t	c	66.61	2.96×10^{-5}	+++-	0.38	ROBO1
3	rs1159913	79193997	a	g	66.62	3.05×10^{-5}	+++-	0.35	ROBO1
3	rs1159912	79194332	a	g	66.62	3.41×10^{-5}	+++-	0.36	ROBO1
3	rs60200150	79191634	a	c	66.74	3.61×10^{-5}	+++-	0.45	ROBO1
3	rs2168373	79193739	c	g	66.62	4.6×10^{-5}	+++-	0.38	ROBO1
3	rs12486635	79197710	a	g	66.61	5.06×10^{-5}	+++-	0.41	ROBO1
3	rs7631406	79190590	t	c	33	5.37×10^{-5}	---+	0.46	ROBO1
3	rs2311351	79197773	t	c	66.61	5.75×10^{-5}	+++-	0.4	ROBO1
3	rs7638301	79196370	c	g	33.28	6.18×10^{-5}	---+	0.4	ROBO1
3	rs10865573	79195497	t	c	33.29	6.4×10^{-5}	---+	0.43	ROBO1
3	rs1378633	79195792	t	c	33.29	7.4×10^{-5}	---+	0.41	ROBO1
3	rs1378632	79195917	t	c	66.71	7.49×10^{-5}	+++-	0.39	ROBO1
3	rs7619949	79190685	c	g	67	8.29×10^{-5}	+++-	0.48	ROBO1
3	rs1383407	78908171	t	c	59.95	8.42×10^{-5}	----	0.96	ROBO1
3	rs6548614	79178697	a	g	70.51	9.16×10^{-5}	+++-	0.35	ROBO1
3	rs9823929	79184872	t	c	29.36	9.19×10^{-5}	---+	0.42	ROBO1
3	rs1378637	79177486	a	g	29.23	1.11×10^{-4}	---+	0.39	ROBO1
3	rs1455832	79176116	a	g	70.53	1.24×10^{-4}	+++-	0.38	ROBO1
3	rs6786179	79177030	a	g	29.29	1.29×10^{-4}	---+	0.38	ROBO1
3	rs1455833	79175867	t	c	29.28	1.47×10^{-4}	---+	0.4	ROBO1
3	rs11127636	78922852	a	c	59.34	2.04×10^{-4}	----	0.98	ROBO1
3	rs4680943	78926431	a	t	59.44	3.32×10^{-4}	----	0.98	ROBO1
6	rs7768291	24656571	a	g	85.03	3.51×10^{-4}	++++	0.67	KIAA0319
3	rs73114798	78927132	t	c	36.34	3.64×10^{-4}	++++	1	ROBO1
6	rs3181234	24655165	t	c	14.8	3.88×10^{-4}	----	0.65	KIAA0319
3	rs12495133	78921520	a	c	36.34	4.1×10^{-4}	++++	1	ROBO1
3	rs34840858	79120773	a	t	28.25	4.16×10^{-4}	---+	0.54	ROBO1
3	rs68030029	79128494	t	g	70.94	4.79×10^{-4}	+++-	0.35	ROBO1
6	rs77272080	24679178	a	c	14.62	5.04×10^{-4}	---+	0.59	KIAA0319
6	rs73392549	24677968	a	g	14.58	5.23×10^{-4}	---+	0.59	KIAA0319
3	rs1455824	79137230	a	g	28.65	5.52×10^{-4}	---+	0.37	ROBO1
6	rs11961837	24681787	t	c	15.18	5.54×10^{-4}	---+	0.59	KIAA0319
3	rs189022	78983297	a	t	47.11	5.55×10^{-4}	++++	0.63	ROBO1
6	rs3756819	24665340	a	c	85.39	5.68×10^{-4}	++++	0.71	KIAA0319
6	rs11962639	24683150	a	g	15.02	5.84×10^{-4}	----	0.54	KIAA0319
6	rs6909884	24676985	a	g	14.58	7.31×10^{-4}	---+	0.62	KIAA0319
3	rs11924366	78914742	t	c	37.3	8.22×10^{-4}	++++	0.97	ROBO1
6	rs2143340	24659071	a	g	85.34	8.35×10^{-4}	++++	0.71	KIAA0319

Table S1. Results of the assessment of 13,827 SNPs in nine candidate genes previously implicated in RD and/or SLI (see *Subjects and Methods* section for a complete list of these genes). Here, only association p-values < 0.001 with **b)** PC1 and **c)** IQ-adjusted PC1 are reported.

[a] Single-base indels were not filtered out from the imputed polymorphisms since they were reliably called in the imputation reference (1000 Genomes, Phase I v3), and were tested for association as they could represent coding frameshift variants of biological interest. [b] Frequency of Allele 1. [c] The direction of effect of Allele1 is reported for datasets in the following order: CLDRC-RD, UK-RD, SLIC, CLDRC-ADHD. [d] Test for the homogeneity of effect sizes across the different datasets ($p \geq 0.05$ indicates homogeneous effects).

S1d)

Trait[a]	CLDRC-RD	UK-RD	SLIC	CLDRC-ADHD
WRead	0.36 (0.259)	0.75 (0.039)	0.15 (0.756)	0.05 (0.855)
WSpell	0.35 (0.294)	0.94 (0.019)	0.44 (0.402)	0.42 (0.137)
PD	0.74 (0.017), 0.66 (0.047)[b]	0.66 (0.106)		0.34 (0.202), 0.37 (0.179)[b]
PA	0.54 (0.105)	0.72 (0.054)[c]		0.1 (0.716)
OC	0.48 (0.124)	0.72 (0.039)		0.39 (0.158)
NWR	0.41 (0.233)		0.16 (0.935)	-0.65 (0.01)
ELS			0.31 (0.416)	
RLS			0.7 (0.037)	

S1e)

Trait[a]	CLDRC-RD	UK-RD	SLIC	CLDRC-ADHD
WRead	-0.11 (0.827)	0.64 (0.088)	0.19 (0.533)	-0.42 (0.19)
WSpell	0.07 (0.867)	0.82 (0.029)	-0.04 (0.885)	-0.07 (0.835)
PD	0.15 (0.731), 0.23 (0.61)[b]	0.83 (0.014)		-0.28 (0.395), -0.41 (0.236)[b]
PA	-0.06 (0.915)	0.77 (0.034)[c]		-0.26 (0.417)
OC	0.56 (0.202)	0.66 (0.031)		0.44 (0.184)
NWR	-0.37 (0.472)		0.71 (0.053)	0.19 (0.514)
ELS			0.77 (0.006)	
RLS			0.7 (0.012)	

S1f)

Trait[a]	CLDRC-RD	UK-RD	SLIC	CLDRC-ADHD
WRead	0.19 (0.556)	0.61 (0.015)	0.08 (0.74)	0.44 (0.293)
WSpell	0.31 (0.322)	0.78 (0.009)	0.13 (0.623)	-0.36 (0.406)
PD	-0.08 (0.8), 0.28 (0.377)[b]	0.65 (0.017)		0.21 (0.611), -0.13 (0.763)[b]
PA	0.53 (0.098)	0.06 (0.813)[c]		0.21 (0.641)
OC	0.23 (0.441)	0.31 (0.28)		-0.14 (0.775)
NWR	0.51 (0.139)		-0.29 (0.322)	0.16 (0.705)
ELS			-0.72 (0.002)	
RLS			-0.16 (0.453)	

S1g)

Trait[a]	CLDRC-RD	UK-RD	SLIC	CLDRC-ADHD
WRead	-0.39 (0.133)	-0.82 (0.008)	-0.7 (0.062)	-0.49 (0.552)
WSpell	-0.21 (0.419)	-0.92 (0.003)	-0.52 (0.17)	-0.24 (0.763)
PD	-0.2 (0.483), -0.59 (0.042)[b]	-0.7 (0.026)		-0.23 (0.795), 0.16 (0.868)[b]
PA	-0.63 (0.024)	-0.32 (0.243)[c]		-0.37 (0.654)
OC	-0.29 (0.273)	-0.83 (0.005)		0.4 (0.615)
NWR	-0.73 (0.009)		-0.23 (0.416)	-0.25 (0.761)
ELS			0.01 (0.984)	
RLS			-0.49 (0.12)	

S1h)

Trait[a]	CLDRC-RD	UK-RD	SLIC	CLDRC-ADHD
WRead	0.41 (0.721)	0.71 (0.1)	0.83 (0.006)	-0.34 (0.317)
WSpell	0.76 (0.461)	0.76 (0.106)	0.75 (0.018)	0.18 (0.567)
PD	0.31 (0.74), 0.45 (0.646)[b]	0.54 (0.184)		-0.22 (0.544), 0.29 (0.416)[b]
PA	0.2 (0.855)	0.83 (0.04)[c]		0.3 (0.372)
OC	0.16 (0.868)	0.48 (0.239)		-0.24 (0.455)
NWR	0.11 (0.909)		0.79 (0.026)	-0.34 (0.36)
ELS			0.56 (0.054)	
RLS			0.46 (0.215)	

Table S1. Effects of the SNPs **d)** rs12495133 (*ROBO1*), **e)** rs4535189 (*ROBO1*), **f)** rs761100 (*KIAA0319*), **g)** rs2143340 (*KIAA0319*) and **h)** rs2875891 (*ATP2C2*) on the single reading and language traits used in constructing PC1. These were computed for each trait as PLINK Multivariate MQFAM loadings and PLINK univariate QFAM association p-values (in brackets). PLINK multivariate loadings refer to minor alleles (A for rs12495133, rs4535189 and rs761100, G for rs2143340 and T for rs2875891).

[a] Legend: WRead = word reading; WSpell = word spelling; PD = phonological decoding; PA = phoneme awareness; OC = orthographic coding; NWR = nonword repetition; ELS/RLS = expressive/receptive language score. [b] Loading on nonword reading and phonological choice (respectively). [c] Although PA had been excluded from the PCA in UK-RD (due to the low number of measures available, see Chapter 2), it was tested in this case to have a term of comparison to the other datasets.

Chapter 5: Investigating the effect of Copy Number Variants on reading and language traits

This chapter is based on:

Gialluisi, A., Visconti, A., Willcutt, E.G., Smith, S.D., Pennington, B.F., Falchi, M., DeFries, J.C., Olson, R.K., Francks, C. & Fisher, S.E. Investigating the effect of Copy Number Variants on reading and language traits. (*in prep*)

Abstract

Reading and language skills are thought to have overlapping genetic contributions, most of which are still unknown. Part of the missing heritability may be caused by Copy Number Variants (CNVs).

In a dataset of children recruited for a history of dyslexia or ADHD and their siblings, we investigated the effect of CNVs on continuous reading and language traits. First we called CNVs using signal intensity data from Illumina OmniExpress array (~723,000 probes). Then we computed the correlation between measures of CNV genomic burden and the first principal component score derived from several reading and language traits, both before and after adjustment for performance IQ. Finally we screened the genome, probe-by-probe, for association with the principal component (PC) scores, through two complementary analyses. The first used CNV calls and tested for association the CNV state at each probe. The second directly tested for association probe intensity data from the array, through FamCNV.

No significant correlation was found between measures of CNV burden and PC scores and no genome-wide significant associations were detected in probe-by-probe screening. Association analysis using CNV calls revealed nominally significant associations (p ~10^{-2}-10^{-3}) within *CNTN4* (contactin 4) and *CTNNA3* (catenin alpha 3). These genes encode cell adhesion molecules with a likely role in neuronal development and have been already implicated in autism and other neurodevelopmental disorders. An assessment of hotspots of neuropsychiatric CNVs revealed a region nominally associated with PC score (p ~0.02-0.04), within *CHRNA7* (cholinergic nicotinic receptor alpha 7), encoding a ligand-gated ion channel mediating fast synaptic transmission. FamCNV analysis detected a region of association (p ~10^{-2}-10^{-4}) within a frequent deletion ~6 kb downstream of *ZNF737* (zinc finger protein 737, uncharacterized protein), which was also observed in the association analysis of CNV calls. This suggests a potential effect of this deletion on reading and language abilities.

Overall these data suggest that CNVs do not underlie a substantial proportion of variance in reading and language skills. Analysis of additional, larger datasets is warranted, to further assess the potential effects that we found and to increase the power to detect CNV effects on reading and language.

Introduction

Reading disability (RD, or developmental dyslexia) and Specific Language Impairment (SLI) are two of the most prevalent neurodevelopmental disorders, with a prevalence of ≈5-8% among school-aged children (Shaywitz et al., 1990; Tomblin et al., 1997). Both RD and SLI are multifactorial disorders with moderate to high heritabilities (30-70%; Barry et al., 2007; Fisher & Defries, 2002), and are characterized by high comorbidity (43-55%; Mcarthur et al., 2000, Snowling et al., 2000), also with other neurodevelopmental disorders such as Attention Deficit Hyperactivity Disorder (ADHD; Pennington, 2006; Willcutt et al., 2010) and Speech Sound Disorders (SSD; Newbury & Monaco, 2010; Pennington & Bishop, 2009). It is likely that RD and SLI share some genetic/neurobiological mechanisms (Newbury et al., 2011; Paracchini, 2011). This has been supported by studies that reported significant genetic correlations between reading and language measures, both in twin (Harlaar et al., 2008) and in family studies (Logan et al., 2011).

Genetic basis of RD and SLI

Genes that have been implicated in RD by linkage and positional/biological candidate approaches include *DYX1C1* (15q21; Nopola-Hemmi et al., 2000; Taipale et al., 2003), *KIAA0319* and *DCDC2* (6p22; Cope et al., 2005; Francks et al., 2004; Meng et al., 2005), *MRPL19/GCFC2* (2p12; Anthoni et al., 2007) and *ROBO1* (3p12; Bates et al., 2011; Hannula-Jouppi et al., 2005). Similarly, some loci have been implicated in SLI, specifically *CNTNAP2* (7q35; Vernes et al., 2008); *CMIP* and *ATP2C2* (16q23-24; Newbury et al., 2009), and *FOXP2* (7q31; Fisher & Scharff, 2009). Some of these genes have roles in important processes in Central Nervous System (CNS) development, such as neuronal migration, axonal guidance and neurite outgrowth (Carrion-Castillo et al., 2013; Vernes et al., 2011; Poelmans et al., 2011). An influence of steroid hormone-related biology on reading and language skills has also been hypothesized, in light of interactions between *DYX1C1* and estrogen receptors (Massinen et al., 2009; Tammimies et al., 2012), and of studies linking sex hormone biology, language performance, and the architecture of brain areas underlying reading and language (Good et al., 2001; Lombardo et al., 2012; Whitehouse et al., 2012). Some of the RD/SLI candidate genes above have shown association with both reading and language measures, suggesting that they may contribute to both RD and SLI (Bates et al., 2011; Newbury et al., 2011; Scerri et al., 2011). In these genes, most of the variants that have

been linked to reading and/or language traits are Single Nucleotide Polymorphisms (SNPs), some of which may also show effects in the general population (Bates et al., 2011; Paracchini et al., 2008; Paracchini, 2011; Scerri et al., 2011; Whitehouse et al., 2011). However, other types of genetic variants have also been implicated in reading and/or language skill variance. These include balanced translocations disrupting *ROBO1* (Hannula-Jouppi et al., 2005) and *DYX1C1* (Nopola-Hemmi et al., 2000; Taipale *et al.*, 2003) in dyslexic families, and translocations and deletions affecting *FOXP2* in a severe form of speech and language delay, called Childhood Apraxia of Speech (CAS) (Fisher & Scharff, 2009).

To help elucidate the shared genetic bases of RD and SLI, three studies have carried out Genome-Wide Association Scans (GWAS) using measures of both reading and language. Although none of them reported genome-wide significant associations, they detected suggestive associations in *ABCC13* (21q11.2) and *DAZAP1* (19p13.3) (Luciano et al., 2013), *ZNF385D* (3p24.3; Eicher *et al.*, 2013), and *FLNC* (7q32.1) and *RBFOX2* (22q12.3) (see Chapter 3; Gialluisi et al., 2014).

CNVs implicated in RD, SLI and related disorders: a brief review

SNP associations reported so far can explain only a small proportion of heritable variance in reading and language skills (Newbury & Monaco, 2010; Peterson & Pennington, 2012). Part of this "missing heritability" may be represented by Copy Number Variants, defined as structural variations in the genome that result in regions larger than 1 kb showing a non-diploid copy number. Several CNVs have been identified in severe neurodevelopmental and neuropsychiatric disorders, including schizophrenia (SCZ), Autism Spectrum Disorders (ASD), Intellectual Disability (ID) and Developmental Delay (DD) (Girirajan et al., 2011; Grayton et al., 2012; Malhotra & Sebat, 2012; Stefansson et al., 2014). However, only a few studies have focused on less severe and pervasive disorders like RD and SLI. In the majority of these studies, a perfect co-segregation between CNVs and RD/SLI status has seldom been observed. In fact, CNVs that are supposed to be pathological are often detected also in phenotypically normal or mildly impaired carriers (Burnside et al., 2011). This complicates the interpretation of associations between CNVs and these disorders.

In a recent investigation on 10 Indian dyslexic families, presenting 14 RD cases and 24 controls, seven de novo CNVs were identified in five cases at different loci, namely *GABARAP* (17p13.1), *NEGR1* (1p31.1), *ACCN1* (17q11.21), *DCDC5* (11p14.1) and the

known SLI candidate *CNTNAP2* (7q35) (Veerappa et al., 2013a). Network analysis of these genes suggested enrichment for interactions in five pathways, including synaptic transmission, axon guidance, transmission of nerve impulse, neurogenesis and cell migration (Veerappa et al., 2013a). In spite of the biological functions of these loci, which make them good candidate susceptibility genes for RD, the lack of evidence of transmission across generations and the low sample size of the study suggest caution in the interpretation of the results. In a parallel study on the same families, focusing on the X chromosome, six dyslexic male subjects in three families were found to carry CNVs disrupting the *PCDH11X* gene (Xq21.31-q21.32; Veerappa et al., 2013b). Although in this case there was more convincing evidence of co-segregation across multiple generations in one family, the CNV events were hypothesized to be generated by distinct and independent unequal recombinations between sex chromosomes at the pseudo-autosomal region Xq21.3 (Veerappa et al., 2013b). In a Dutch family, Poelmans and colleagues (2009) identified a heterozygous deletion in 21q22.3, co-segregating with RD in the father and his three sons. This deletion, which spanned ~176 kb, encompassed four genes, namely *PCNT*, *DIP2A*, *S100B*, and *PRMT2*, and haploinsufficiency of one or more of these genes was hypothesized to contribute to RD susceptibility (Poelmans et al., 2009).

Other CNVs have been associated with poor reading performance in the context of other comorbid disorders. Pagnamenta and colleagues (2010) found out that a ~600 kb deletion disrupting both *DOCK4* and *IMMP2L* (7q31.1) -previously identified in two brothers affected by ASD- was co-segregating with poor reading performance in six out of nine relatives of the affected siblings (Pagnamenta et al., 2010). Another *DOCK4* exonic deletion, co-segregating with the dyslexic status, was found in a distinct family where both the father and the son presented with RD (Pagnamenta et al., 2010).

The largest study to date on CNVs in dyslexia, involving 376 RD cases, together with 350 ASD cases, 501 ID cases, and 337 controls, reported various candidate susceptibility CNVs. These included one heterozygous deletion disrupting *IMMP2L*, co-segregating with dyslexia in a family, and two distinct deletions overlapping *AUTS2* (7q11.22; Girirajan et al., 2011), a well-known ASD susceptibility locus (Oksenberg & Ahituv, 2013; Sultana et al., 2002). One of these *AUTS2* CNVs showed imperfect co-segregation with RD. This study also reported significant associations of the genomic burden of rare large CNVs with ASD and ID, but not with RD (Girirajan et al., 2011).

Candidate susceptibility genes were recently proposed also for SLI, based on CNV analysis. Wisznieski et al. (2013) identified a heterozygous deletion on 2q36, co-segregating with language delay and white matter hyper-intensities in 15 Southeast Asian families. This deletion disrupted the gene *TM4SF20*, and appeared to be a founder mutation in Southeast Asian populations (Wisznieski et al.; 2013). In a family-based cohort including SLI cases and their first degree relatives, a ~21 kb exonic microdeletion within *ZNF277* (7q31.1, adjacent to the *IMMPL2/DOCK4* locus) was found in homozygous state in an affected girl and in heterozygous state in her parents, both with histories of language problems (Ceroni et al., 2014). In spite of an increased frequency of this CNV in SLI cases compared to controls (1.1 vs 0.4%), there was incomplete segregation with SLI in this and other families. The deletion was not inherited by SLI probands in some families, and was inherited by unaffected siblings in other families (Ceroni et al., 2014). More recently, a genome-wide CNV study comparing 127 independent SLI cases from the same dataset, together with 385 first-degree relatives and 269 unrelated controls, reported candidate de novo CNVs in SLI cases. These disrupted the genes *ACTR2* (2p14), *CSNK1A1* (5q33.1) and the regions typically involved in 22q11.2 and 8p23.1 duplication syndromes. Pathway analysis of the CNVs detected in SLI cases revealed a significant overrepresentation of genes related to acetylcholine binding, cyclic-nucleotide phosphodiesterase activity and MHC proteins. Interestingly, this study also reported a significant difference in the genomic burden of CNVs between cases and unrelated controls (Simpson et al., 2015).

CNVs have been associated with poor language performance also in the context of other neuropsychiatric disorders: Raca et al. (2013) reported two patients with 16p11.2 microdeletion syndrome -characterized by mild cognitive impairment, general developmental delay, speech and language delays and autistic disorder- meeting criteria for CAS. Newbury and colleagues (2013) later reported a subject with CAS and pervasive developmental disorder, where a similar 16p11.2 *de novo* deletion was hypothesized to act jointly with an inherited 6q22.31 duplication. Similarly, patients presenting deletions/duplications of the 15q11.2 region (BP1-BP2), typically involved in Prader-Willi/Angelman syndrome, were reported to frequently exhibit speech and language delays (Burnside et al., 2011). Another interesting candidate for CAS and SSD derived from the comparison of several subjects with 2p15-p16.1 microdeletion syndrome, a disorder characterized by a broad phenotypic spectrum including cognitive, linguistic and psychiatric disabilities. The report of a patient carrying a heterozygous de novo deletion encompassing a single gene, *BCL11A* (2p16.1), and

with a mild phenotype characterized by apraxia, dysarthria and expressive language delay, led to hypothesize that this gene may be a susceptibility locus for these disorders (Peter et al., 2014).

Recently, Stefansson and colleagues (2014) investigated the effect of several CNVs previously associated with schizophrenia or autism (hereafter called "neuropsychiatric CNVs") on different cognitive traits in a big sample of the Icelandic population (N~102,000). By comparing SCZ patients, neuropsychiatric CNVs carriers, other CNVs carriers and general population (non-carriers) controls, they found that neuropsychiatric CNVs carriers performed at a level between SCZ patients and general population controls on several psychometric tests, suggesting an effect of these CNVs on general cognition. Six neuropsychiatric CNVs were associated with verbal or performance IQ, namely 16p11.2 deletion and the reciprocal duplication, 17p12 deletion, 17q12 duplication, 16p12.1 deletion and 16p13.1 duplication. Focusing on the cognitive traits most relevant to language and reading, 16p11.2del and 22q11.21dup were associated with category and letter fluency, while 15q11.2del was associated with history of dyslexia and dyscalculia, although the associations were weakened after conditioning on IQ. Interestingly, CNVs in the 15q11.2 region also showed an allele-dose-dependent effect on structural brain measures related to reading and language: deletion carriers exhibited a bilateral reduction of white matter in the temporal lobe and an increase in the volume of corpus callosum (Stefansson et al., 2014).

The present study

In the present work, we investigated in detail the potential influence of CNVs on reading and language skills. To do so, we used one of the datasets already involved in our previous GWAS meta-analysis (GWASMA, Chapter 3), composed of children recruited for school history of RD or ADHD, and their unaffected siblings. Through a comprehensive approach, we first investigated the effect of CNVs on a categorical definition of RD in probands and their siblings. Then we tested CNVs for association with continuous reading and language traits, through a first Principal Component score (PC1) representing a substantial proportion of the shared variance in these traits. To detect genetic effects on PC1 largely or wholly independent of general cognitive abilities, an IQ-adjusted version of the PC1 score (IQ-adjusted PC1) was also computed and tested for association along with the original PC1

score. Two complementary genome-wide strategies were used to test association between CNVs and continuous reading/language traits.

First we analysed the association between the CNV state at each probe in the DNA array and our principal component (PC) scores, considering both deletions and duplications at each location as a single 'CNV-positive' (CNV+) state, in contrast to the diploid state. This was aimed at detecting effects of CNVs irrespective of the abnormal copy number, in other words assuming that either deletion or duplication would impact in the same direction and to the same extent on cognitive performance.

Second, we analysed the association between raw intensity data for each probe and the PC scores, to detect dosage-dependent effects of common multi-allelic CNVs. This approach was complementary to the association test with CNV state for two main reasons: first, because it tested directly raw intensity data, reducing the loss of information and of power implied by translating these continuous data into discrete copy number states; second, because it aimed at detecting dosage-dependent effects of common multi-allelic CNVs (or Copy Number Polymorphisms, CNPs) on the continuous traits tested (Eleftherohorinou et al., 2011; Falchi et al., 2014).

Current CNV research in psychiatric genetics (reviewed in the paragraph above) often relies on case/control dichotomous classifications and seldom detect perfect co-segregation between CNVs and disease status (Burnside et al., 2011). This may overcomplicate the interpretation of results and the postulation of causality links. When heritable quantitative traits are available that are strongly correlated with a dichotomous definition of a disorder -as in the case of reading/language traits and RD/SLI- analyzing the effect of putative CNVs directly on the quantitative trait provides an effective alternative to the analysis of co-segregation between CNVs and the disorder. The former analysis is aimed at detecting variants with reduced penetrance and variable expressivity on our traits of interest, while the latter one is aimed at detecting variants with full penetrance and expressivity. As these approaches are complementary, we decided to use both in our study.

Subjects and Methods

The experimental workflow of the present study, described in this section, is summarized in Figure 1. For simplicity, genotype and phenotype QC are described below in single paragraphs.

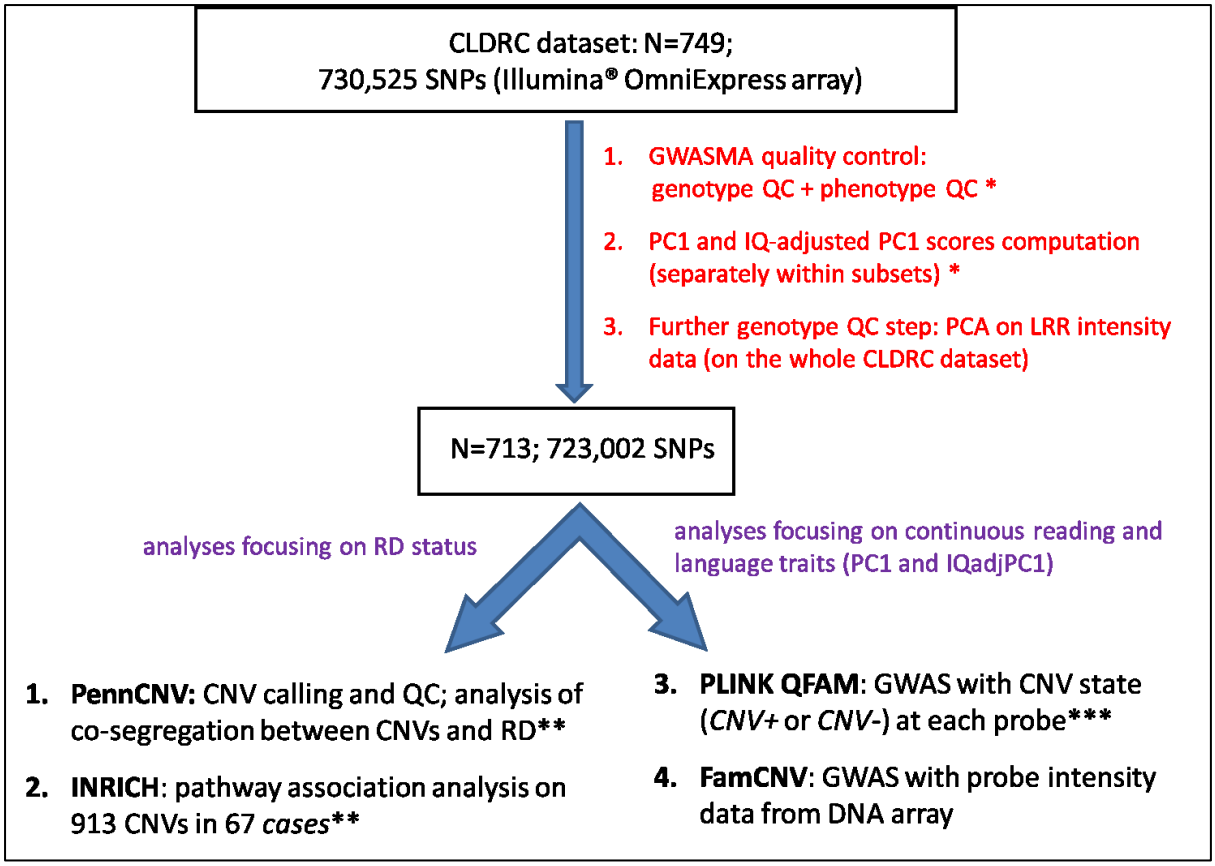


Figure 1. Experimental workflow and dataset analyzed in the present study. *As described in Chapters 2 and 3. ** RD cases were defined as samples in the lowest 10% of IBG discriminant score distribution. *** Legend of CNV states: "CNV+" corresponds to copyN ≠2 (≠1 for X chromosome probes in males); "CNV-" corresponds to copyN =2 (=1 for X chromosome probes in males). See paragraph *Genome-Wide Association Scan (GWAS) with CNV state* for further details.

Dataset

The dataset analyzed in the present work was collected in the Colorado Learning Disabilities Research Centre (CLDRC) study, an ongoing research on the etiology of learning disabilities carried out in 27 school districts in Colorado, USA (Defries et al., 1997; Willcutt et al., 2005). This dataset was recently analysed in a GWAS meta-analysis of reading and language traits (Gialluisi et al., 2014; Chapter 3). Briefly, pairs of twins were recruited for a school report of RD or ADHD in at least one of the twins; they were then administered a number of psychometric tests for several learning-related skills, along with their additional co-siblings, and DNA was collected for genetic studies. The Institutional Review Boards of the University of Nebraska Medical Center and of the University of Colorado at Boulder had approved the protocol, and written informed consent of the participants (or their parents) was obtained.

For the present study, for MZ twin pairs, we selected one child per pair based on the maximum availability of reading- and language-related trait data, or otherwise randomly. The sample of twins and siblings available for this study comprised 749 participants in total (mean age 11.7 years, age range 8-19), from 343 unrelated twinships/sibships. Of these, 266 of the twinships/sibships (a total of 585 participants) were originally recruited via a proband with a history of RD, and 77 of the twinships/sibships (164 participants in total) were originally recruited via a proband with a history of ADHD. The two subsets are indicated hereafter as CLDRC-RD and CLDRC-ADHD. The other datasets originally involved in our GWAS meta-analysis (Gialluisi et al., 2014; Chapter 3) -namely SLIC and UK-RD- were not included in the present study as CNV analysis of these datasets has been already published elsewhere (for SLIC; Ceroni et al., 2014; Simpson et al., 2015) or planned by other groups (for UK-RD).

Reading and language measures

The reading- and language-related traits assessed in CLDRC are reported in Table1 and the relevant measures are described in detail in Chapter 2 (Table S1c). These traits had been previously age-adjusted according to normative data (Compton et al., 2001; Friend & Olson, 2010) and further rank-normalized when a measure differed significantly from normality. Phenotypic outliers were removed from the dataset, along with subjects with full scale IQ < 70 (two participants in CLDRC-RD in total). This left 564 subjects in CLDRC-RD and 163 in CLDRC-ADHD. Then samples underwent separate Principal Component Analyses (PCAs) in CLDRC-RD and in CLDRC-ADHD for the computation of the First Principal Component scores within each dataset, as briefly described below (further details in Chapter 2).

First Principal Component score

The First Principal Component (PC1) from all of the language- and reading-related traits available (Table 1) was derived in each dataset, through the SPSS® 20.0 Factor Analysis. Only linear components with Eigenvalue>1 were extracted, allowing for correlation among the components (oblique rotation) and excluding subjects with any missing measure. PC1 explained 64.5% of the total variance in CLDRC-RD and 52% in CLDRC-ADHD, while PC2 explained no more than 13% of the total variance in both datasets. PC1 scores showed a broad pattern of loadings across the traits in both datasets (Table 1). To get a measure of

common variance in reading and language skills independent of general cognitive abilities, we also regressed PC1 against performance IQ (which had not been included in PC1 computation), again separately within the two datasets, and used the residuals as IQ-adjusted PC1 scores (IQadjPC1).

Trait[a]	Description (ability assessed)	CLDRC-RD (564)	CLDRC-ADHD (163)
WRead	Reading real words	0.918	0.871
WSpell	Spelling real words	0.813	0.764
PD	Ability to convert letter strings into sounds, according to given phonetic rules	0.895, 0.861[b]	0.821, 0.729[b]
PA	Ability to recognize and manipulate speech sounds (phonemes)	0.801	0.744
OC	Ability to recognize a word as an orthographic unit and to retrieve the corresponding phonological form	0.764	0.644
NWR	Ability to repeat nonsense words orally presented	0.493	0.355
VIQ	Verbal reasoning		
PIQ	Logical reasoning		
PC1	Common variance in reading and language skills	544 (528)	159 (155)
IQ-adjusted PC1	Common variance in reading and language skills, not shared with general cognitive abilities	544 (528)	159 (155)

Table 1. Phenotypic traits available and measures used for PC1 score derivation (labelled with relative loadings on PC1, as already shown in Table 1 of Chapter 2). Sample sizes of the datasets that underwent the PCA are reported in the header row. Number of samples for which PC1 score was computed are reported at the bottom of the table (as we excluded participants with at least one missing measure among the traits involved in the PCA). These numbers still include LRR outliers and samples discarded in CNV calling and QC process, which were filtered out for the specific purpose of this study, after extraction of PC1 scores. Final sample sizes at the end of all QCs are reported in brackets. [a] Legend: WRead = word reading; WSpell = word spelling; PD = phonological decoding; PA = phoneme awareness; OC = orthographic coding; NWR = nonword repetition; ELS/RLS = expressive/receptive language score; VIQ/PIQ = verbal/performance IQ. [b] Loadings of nonword reading and phonological choice (respectively) on PC1s.

IBG discriminant score

For this study, we used an additional phenotypic trait, the IBG discriminant score (called IBGdiscr hereafter), a discriminant function empirically developed to diagnose RD in the context of the CLDRC study (Defries, 1985). This score is a composite measure of word recognition, spelling and comprehension subtests of the Peabody Individual Achievement Test (Dunn & Markwardt, 1970; further details available in *Supplementary Material S1*). For the purpose of this study, we used IBGdiscr to select all the participants in the first decile of the score distribution (Figure S1a), namely all the subjects with a standardized IBGdiscr < -1.4 (N = 67), as representative of poor reading performance. For simplicity, we will indicate these subjects as "RD cases" in the analyses where a dichotomous case-control classification will be needed (see below).

Pairwise trait correlations of the reading and language composite/component scores analyzed -computed as median Pearson's r coefficients over 100 repeat random samplings of one individual from each unrelated sibship- were high (r~0.83-0.98), both in CLDRC-RD and in CLDRC-ADHD (see Table S1).

DNA array data: generation and quality control

The two subsets were treated as a single dataset in DNA data generation and quality control (QC), as previously described in our GWAS meta-analysis (Chapter 3). DNA was extracted from whole blood or buccal swab samples and prepared for genotyping using standard protocols. DNA array data were generated using Illumina® Human OmniExpress array (730k SNPs) and data were processed using Illumina's GenomeStudio® software, following the manufacturer's guidelines. QC and CNV calling process (see below) followed procedures already used in previous CNV studies (Elia et al., 2012; Szatkiewicz et al., 2013; Simpson et al., 2015). Samples with genotyping success rate < 95% were discarded in GenomeStudio, along with probes mapping as "0" (no position) and "Y" (Y chromosome) and probes with call frequency < 95%. Using functions in the software PLINK v1.07 (Purcell et al., 2007; http://pngu.mgh.harvard.edu/~purcell/plink/), we filtered out samples which showed inconsistencies in genome-wide identity-by-descent sharing with their siblings and unrelated individuals, or sex mismatches, or call rates < 98%, as well as homozygosity outliers, as described in Chapter 3 (see *Supplementary Material S1* for details).

As a further QC step for this study, we ran a PCA on the Log R Ratio (LRR) intensity signals of the 723,002 probes passing QC, through the *pca* command (*singular value decomposition* method) in the *pcaMethod* R package (Stacklies et al., 2007; R Core Team, 2013), extracting the first 100 principal components. This allowed to assess the absence of extreme batch effects among the different plates of the microarray and to detect and remove 14 LRR outliers (Figure S1d), which left 713 subjects for subsequent analysis.

Copy Number Variants (CNVs) calls

To detect CNVs, we applied PennCNV (version June 2011, Wang et al., 2007) separately for autosomes and the X chromosome (704,855 and 18,147 SNPs, respectively), analyzing the two subsets jointly (N=713). PennCNV is a Hidden Markov Model (HMM) based algorithm designed for Illumina® platforms, which makes use of normalized intensity data -both Log R

Ratio (LRR) and B Allele Frequencies (BAF)- of probes to call putative CNVs in SNP microarrays (Wang et al., 2007). For this analysis, we built a custom Population B allele Frequency (PBF) file from our array intensity data through the *compile_pfb.pl* script in PennCNV, while default HMM parameters and GC model signal adjustment file (to reduce false positive calls) were used. In order to obtain highly reliable CNV calls, we applied a series of filters to the CNV events initially called through the *detect_cnv.pl* script: only putative CNVs with a minimum confidence score of 10, covering at least 10 kb and 3 consecutive SNPs and showing limited overlap (<50%) with Ig regions, pseudo-autosomal regions (PARs), centromeres or other large genomic gaps were selected. Moreover, to ensure only high quality of samples, we filtered out samples showing an excessive number of CNV calls (>100 autosomal CNVs per sample) and LRR standard deviation > 0.35. All the other parameters for samples filtering were set to default. Close CNVs were joined when the gap separating them was ≤20% of the total length of the region that they covered. CNVs passing QC were finally annotated to RefSeq genes (including both protein coding and non-coding sequences, such as microRNAs), within 50 kb beyond the 5'- and 3'-untranslated regions (UTRs), to include CNVs overlapping potential regulatory regions. At the end of this process we had 10,110 final CNV calls for 705 samples, of which 6,627 were annotated to genes.

Interpretation of CNVs and general statistics

The samples passing PennCNV QC (N=705) were tested for correlation between their CNV burden -both in terms of total length and of total number of CNV events per sample- and our continuous traits of interest, namely PC1 and IQadjPC1, separately in the two subsets. This analysis was applied to 528 PC1/IQ-adjusted PC1 measures available in CLDRC-RD and to 155 measures available in CLDRC-ADHD. To generate correlations unbiased by non-normality of CNV burden measures and by sample relatedness, rho correlation coefficients were calculated as the median rho over 100 repeat random samplings of one individual from each unrelated sibship, in R (http://www.r-project.org/; R Core Team, 2013).

The final annotated CNVs were also assessed individually for co-segregation with the "RD case" status as defined above, focusing on large CNVs (>500 kb), on CNVs shared between two or more affected co-siblings, and on CNVs disrupting particular RD/SLI candidate genes (reviewed in the *Introduction* section) or overlapping with other neuropsychiatric CNVs (previously tested by Stefansson et al., 2014).

Genome-wide CNV association analyses of continuous reading and language PC traits

Genome-Wide Association Scan (GWAS) with CNV state

CNV calls made in PennCNV were also used for a genome-wide association test between CNV state at each probe and PC1/IQ-adjusted PC1. The alternative CNV states at each probe were "CNV-negative" (CNV-) when a probe showed a diploid copy number, and "CNV-positive" (CNV+) when it showed an abnormal copy number. In other words, both deletions and duplications at each probe were considered as a single CNV+ state.

We applied PLINK v1.07 QFAM analysis (Purcell et al., 2007) to all the 50,825 probes covered by CNV events (48,702 autosomal probes and 2,123 X chromosome probes), in CLDRC-RD (N=528) and CLDRC-ADHD (N=155) separately. This method is normally used in association analyses of family-based datasets, to test for association at each SNP by regressing trait scores on genotypes in an additive linear model.

In order to have a bi-allelic coding for probes involved in this analysis, which indicated the presence or absence of a non-diploid state, fake genotypes were created in the .ped input files. These were coded as "11" when the probes were not covered by any CNV (i.e. copy number =2) and as "12" when they fell within CNV calls (i.e. copy number ≠2). For chromosome X, CNV states per probe were coded as "11" for probes with copy number =1 and "12" for probes with copy number ≠1 in males, while they were coded following the rules of autosomal CNV state in females. Then X chromosome probes were tested for association separately within males and females, and later meta-analysed. To correct for non-independence of siblings, permutations (i.e. label-swapping of phenotypes/genotypes) were run in QFAM analysis (see *Supplementary Material S1* in Chapter 3 for details). After QFAM analysis, the results of separate GWAS for CLDRC-RD and CLDRC-ADHD were meta-analysed in the software METAL (http://www.sph.umich.edu/csg/abecasis/Metal/index.html; Willer et al., 2010), through the sample-size based scheme. This algorithm is normally used to meta-analyse SNP associations and does not assume equivalence of allelic effect sizes between datasets, as described in our GWAS meta-analysis (Chapter 3). In this case, we used it to meta-analyse associations with CNV state at each probe in a genome-wide context. Results were then interpreted in terms of consecutive probes showing significant associations (i.e. at least two consecutive probes with $p < 0.005$ at the genome-wide level), representing regions of overlap of two or more CNVs with potential effects on the continuous traits investigated.

Genome-Wide Association Scan (GWAS) with intensity data

As a complementary analysis, we also tested association of LRR and BAF intensity data from our DNA array with PC1 and IQ-adjusted PC1, applying FamCNV 2.0 (beta version available upon request to Dr. Mario Falchi; Eleftherohorinou et al., 2011). This software tests association between raw intensity data at each probe and continuous traits, taking into account family relations, in a linear mixed effects model where IBD sharing and phenotypic covariance are treated as random components.

In this analysis we tested for association with PC1/IQadjPC1 704,855 autosomal probes passing QC in CLDRC-RD (N=528) and in CLDRC-ADHD (N=155), using as covariates the first and second principal components computed in the PCA of LRR data (see *DNA array data: generation and quality control* paragraph). After running separate GWAS in the two subsets, the results were meta-analyzed as above, using rho correlation coefficients between LRR data and PC1/IQ-adjusted PC1 as beta values at each probe, indicative of the direction of association. Results were interpreted in terms of contiguous probes showing significant associations (i.e. pairs of consecutive probes with p-value < 0.001 and contiguous with two or more probes with p < 0. 05), which were more likely to represent real CNV events.

Pathway-based analysis of CNV calls

To test specific molecular networks for enrichment of potentially disrupting CNVs, we ran a pathway-based association analysis in INRICH v1.0 (Lee et al., 2012; http://atgu.mgh.harvard.edu/inrich/started.html). This tool tests groups of independent associated genomic intervals for enrichment of overlaps with predefined gene sets, through a permutation-based approach. For the present analysis, we defined as associated genomic intervals 913 CNVs called in 67 RD cases (resulting in 259 non-overlapping intervals). For each candidate gene set, INRICH counted the number of target genes which overlapped with at least one interval, through the *TARGET* algorithm (specifically designed for CNV analysis; Lee et al., 2012). Gene boundaries were again defined as extending 50 kb beyond the 5'- and 3'-UTRs, while random genomic intervals simulated in the permutations of the test were extracted from a reduced set of 50,825 SNPs, namely all the probes encompassed by CNV calls. We considered testing CNV calls more suitable than testing associated genomic intervals as produced by GWAS analyses, since such intervals would need to be defined on an

LD basis, which is clearly inappropriate for the analysis of CNVs. An approach comparable to ours has already been used in a CNV study of SLI cases (Simpson et al., 2015).

Initially we tested for enrichment three candidate gene lists, based on the gene sets of the Gene Ontology Database (http://www.geneontology.org/) and representing three distinct neurobiological hypotheses on the etiology of reading and language disabilities: axon guidance (including all the GO sets containing the term "axon guidance"), neuronal migration (including all the GO sets containing the term "neuron migration"), and sex hormones biology (including all the GO sets containing the terms "steroid", "androgen", "estrogen", "progesterone" and "testosterone"). These candidates had been already tested in our previous GWAS meta-analysis (Chapter 3), where more detailed explanation on the evidence leading to test these pathways and on the specific parameters used is available. Then, we extended the assessment to 1748 GO sets containing at least 10 genes, for exploratory purposes.

Results

CNV calls

General CNV burden statistics

After QC, there were 10,110 final CNV calls, of which 6,627 were annotated to genes within a 50 kb interval from the UTRs, for 705 samples. These showed a median number of 13 CNVs per sample (9 considering only annotated CNVs), and a median total length of ~916 kb covered by CNVs per sample (~681 kb considering only annotated CNVs).

Correlation assessments between CNV burden measures (both CNV number and total length) and our continuous traits of interest -PC1 and IQadjPC1- did not reveal any significant correlation in the two CLDRC subsets (maximum correlation rho~ -0.1, p= 0.37).

Large CNVs

Large CNVs are more likely to span multiple genes and to have deleterious effects than smaller CNVs (Girirajan et al., 2011). Among CNVs spanning more than 500 kb in RD cases (Table 2), a heterozygous duplication was detected in two affected siblings, but not in their unaffected co-sibling (with IBGdiscr= -0.62, PC1= -0.47 and IQadjPC1= -0.42). This large CNV spanned ~1.2 Mb in the pericentromeric region 11q11-q12.1, covering several *OR* genes (encoding olfactory receptors) and *TRIM* genes (encoding tripartite motif proteins).

CNVs shared between RD cases

Among all the sibships analysed, ten presented more than one RD case. In these sibships we assessed annotated CNVs which were shared between two or more affected co-siblings, regardless of their length. We investigated these variants as they were more likely to confer genetic susceptibility to reading impairment, compared to CNVs presented by single cases. A total of six CNV events fell in this category (Table 3), including the large duplication mentioned above and other five CNV events, described below.

Two heterozygous duplications in 18q11.2 showed a partial overlap of ~28 kb, encompassing 5'-UTR and exon 1 of *ZNF521* (zinc finger protein 521). This overlap was detected between two affected siblings in a unique RD family presenting three cases. We detected another heterozygous duplication in a downstream intronic region of *ZNF521* (data not shown), in an unrelated unimpaired participant (IBGdiscr= -0.27, PC1= -0.99, IQadjPC1= -0.75).

A shared heterozygous duplication, spanning ~96 kb on 13q32.1, overlapping with 5'-UTR and exon 1 of the gene *ABCC4* (ATP-binding cassette, subfamily C, 4), was observed in two cases in an RD family, but not in their unaffected sibling. This duplication was also found in other three unrelated subjects: two of them were the worst performing siblings in their families (IBGdiscr -0.87 and -0.21, PC1 -0.35 and -0.01, IQadjPC1 0.22 and 0.92) and one was a singleton with normal reading and language skills (IBGdiscr= 1.66, PC1= -0.19, IQadjPC1= -0.12).

In another family, presenting two affected siblings but no unaffected co-siblings, we detected two shared CNVs (both heterozygous duplications), which were not detected in any other participant in the study. One of them, spanning ~27 kb on 6q24.2, covered the last 9 exons (66-74) in the 3' terminal region of the *UTRN* (utrophin) gene, including its 3'-UTR. The other one spanned for ~63 kb and overlapped exons 37-48 within *DNAH14* (dynein axonemal heavy chain 14) on 1q42.12.

Finally, we observed a small (~10 kb) heterozygous deletion overlapping *MIR5684* (microRNA 5684, 4q32.3) in two cases of a family presenting no other siblings.

CNVs in genes previously associated with RD, SLI and correlated traits

We identified seven putative CNVs annotated to candidate susceptibility genes that have been implicated in RD/SLI by more than one study (Anthoni et al., 2007; Bates et al., 2011; Cope

et al., 2005; Fisher & Scharff, 2009; Francks et al., 2004; Hannula-Jouppi et al., 2005; Meng et al., 2005; Newbury et al., 2009; Nopola-Hemmi et al., 2000; Taipale et al., 2003; Vernes et al., 2008). These CNVs are reported in Table S2a. Among the candidate RD/SLI genes, *ROBO1*, *DYX1C1*, and *CNTNAP2* were overlapped by one or more of these CNVs. However, only two of the seven participants showing these variants were impaired and none of these CNVs co-segregated with poor reading-language performance (Table S2a).

Similarly, we detected seven CNV calls overlapping genes in which suggestive associations were detected in previous GWAS studies of reading and language skills (Eicher et al., 2013; Luciano et al., 2013; Gialluisi et al., 2014). A list of these CNVs is reported in Table S2b. Again, none of these variants co-segregated with RD status or with poor reading-language performance.

CNVs previously implicated in RD/SLI and common neuropsychiatric CNVs

We checked our CNV calls for overlaps with genes and regions previously found to be disrupted by CNVs in subjects with RD/SLI or weak reading/language performance (Ceroni et al., 2014; Girirajan et al., 2011; Pagnamenta et al., 2010; Peter et al., 2014; Poelmans et al., 2009; Veerappa et al., 2013a; 2013b, Simpson et al., 2015; Wiszniewski et al., 2013; reviewed in the *Introduction* section). Table S2c reports these CNVs, which were detected in *NEGR1*, *IMMP2L*, *PCDH11X*, *CNTNAP2, CSNK1A1, MSRA* (8p23.1 region), *UBASH3B, CACNA2D1, GPC5, VWA3B, CXorf22, TM4SF20* and in several genes in the 22q11.21 region. Again, none of these variants showed co-segregation with poor reading-language performance in the sibships.

Similarly, we assessed overlaps with common neuropsychiatric CNVs (typically involved in autism and schizophrenia) recently assessed by Stefansson and colleagues (2014) for effects on several cognitive traits in a large sample of the Icelandic population. Table S2d reports a list of canonical CNVs detected in our study (i.e. largely or completely overlapping the above mentioned neuropsychiatric CNVs, reported in Table S1 in Stefansson et al., 2014). Among these CNV events, a 1.33 Mb heterozygous duplication in 16p13.11 was detected in an affected participant, who was the worst performer in his sibship and exhibited strong score discrepancies with his co-sibling (>3 for IBGdiscr and PC1 and >2.6 for IQadjPC1). However, a similar duplication was present in an unrelated participant showing normal performance and PC1 and IQadjPC1 scores higher than his sibling (data not shown).

Several other CNVs, showing limited overlap with the neuropsychiatric CNVs assessed above, were detected but are not reported here for space limits (available upon request). When two or more CNV calls were overlapping in these regions, the probes encompassed were assessed in PLINK QFAM analysis of CNV state, to detect stretches of consecutive probes associated with PC1 and IQ-adjusted PC1 scores.

Family-based GWAS of Principal Component scores

Association test with CNV state at each probe

GWAS meta-analysis testing association between CNV state at each probe and PC1/IQ-adjusted PC1 did not report any significant association surviving correction for multiple testing of 6,586 autosomal probes meta-analysed ($\alpha = 7.6x10^{-6}$), namely all the probes encompassed by at least one putative CNV event in both our subsets. None of the 2,123 X chromosome probes lied within CNV events detected in participants of both sexes and in both CLDRC subsets; therefore none of these probes was meta-analyzed. The results of this analysis on an individual probe basis are reported in Tables S3a, b.

These results were interpreted in terms of consecutive probes showing significant associations with PC1 and/or IQ-adjusted PC1 (i.e. at least two consecutive probes with $p < 0.005$), in regions of overlap of two or more CNVs in our dataset (Table 4). All of the top associated regions showed nominally significant associations both with PC1 and IQ-adjusted PC1, with the exception of chr3:2,663,757-2,675,189 and chr11:55,241,556-55,362,955 (p-values ~[0.1-0.12] and ~[0.05-0.1] in PC1 meta-analysis; Table 4). Three of these regions were overlapping genes, namely chr3:2,663,757-2,675,189 (p-values in the range $[3.0; 6.1]x10^{-3}$), lying within *CNTN4* (contactin 4, 3p26.3; Figure 2a); chr10:68,223,696-68,242,672 (p-values ~ $[4.0; 4.3]x10^{-3}$), within *CTNNA3* (catenin alpha 3, 10q21.3; Figure 2b); and chr11:55,241,556-55,362,955 (p-values ~ $[2.0; 4.8]x10^{-3}$), falling in the 11q11 region and encompassing genes *OR4C15* and *OR4C16* (olfactory receptors 15 and 16, family 4, subfamily C; Figure 2c). Frequency of CNV+ state in the top associated regions ranged between 0.6 and 3.0%.

We also checked the presence of nominally significant associations (i.e. at least two consecutive probes with $p < 0.05$) in the regions disrupted by CNVs in RD, SLI or more severe neuropsychiatric disorders (Ceroni et al., 2014; Girirajan et al., 2011; Pagnamenta et al., 2010; Peter et al., 2014; Veerappa et al., 2013a; 2013b, Simpson et al., 2015;

Wiszniewski et al., 2013; Stefansson et al., 2014; reviewed in the *Introduction* section). If CNV events in any of these regions had been called only in one of the subsets and therefore meta-analysis had not been run for the probes encompassed, we assessed directly the GWAS results in the relevant subset. Among the candidate CNVs assessed, a ~164 kb region (chr15:32,350,775-32,514,341) partially overlapping *CHRNA7* (cholinergic nicotinic receptor alpha 7, 15q13.3; Figure 2d) showed a series of 27 consecutive probes (from rs11637923 to rs2611583) associated with PC1 (p-values ~ [0.025; 0.049]). These associations were detected in CLDRC-RD as no CNVs were called in the CLDRC-ADHD subset, and were not significant in the IQ-adjusted PC1 GWAS (p-values ~ [0.055; 0.1]). This region showed a frequency of CNV+ state of ~1.5-2% (see Table S2e for relevant CNV calls) and a positive allelic trend between the CNV+ state and PC1/IQ-adjusted PC1.

Association test with probe intensity data

GWAS meta-analysis of PC1/IQ-adjusted PC1 scores with intensity data (FamCNV), did not reveal any genome-wide significant association surviving correction for multiple testing of 704,855 autosomal probes ($\alpha = 7.1x10^{-8}$). The results of this analysis on an individual probe basis are reported in Tables S3c, d.

Also for this analysis, we were interested in detecting two or more consecutive probes showing significant association. For this purpose, we filtered our association results to detect all the pairs of consecutive probes with p-value <0.001 and contiguous with two or more probes with p <0.05. Such criteria were set to reduce the probability to observe spurious associations due to noise intrinsic to raw intensity data. Although we did not find any region meeting these criteria in the results of the meta-analysis, we found such a region in the GWAS in CLDRC-RD. This ~58 kb region (chr19: 20,657,781 - 20,715,228) consisted of 8 consecutive SNPs on 19p12, associated with both PC1 (top consecutive hits rs2021399 and rs2545918, $p = 9x10^{-4}$ and $5x10^{-4}$ respectively) and IQ-adjusted PC1 ($p = 3x10^{-4}$ and $9x10^{-4}$; see Table S3e). This region lay within a ~80 kb deletion very frequent in our dataset (called in 11.3 % of CLDRC participants, for a total of 80 CNV calls, reported in Table S2f) and ~6 kb downstream of *ZNF737* (zinc finger protein 737, Figure 3). The same region of overlap also showed nominally significant association (p-values ~ [0.01; 0.02]) in the PLINK QFAM analysis with CNV state (paragraph above), in a wider interval (chr19:20,626,179-20,715,228, see Figure 3 and Table S3f). Both in the association test with SNP intensity data

and in the association test with CNV state, this deletion showed a positive effect on PC1/IQ-adjusted PC1.

Pathway-based analysis of CNV calls

Pathway association analysis of 913 CNV calls presented by 67 RD cases did not reveal any significant enrichment surviving correction for multiple testing, neither in the analysis of three composite candidate pathways representing neuronal migration, axonal guidance and steroids-related processes (corrected p-values ~ 1, Table S3g), nor in an exploratory analysis at the pathway-wide level (data not shown). The GO terms showing nominally significant enrichment in the pathway-wide analysis were *carbohydrate binding* ($p = 0.02$), *hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides* ($p = 0.032$) and *rRNA binding* ($p = 0.049$).

Subject	Family	Chr	Start (kb)	End (kb)	SNPs	Length (kb)	CopyN	Gene	PC1	IQadjPC1	IBGdiscr
IBG143157	3914	2	96,196	96,737	26	541	3	FAHD2CP,GPAT2,LINC00342,TRIM43	-2.51	-2.59	-3.29
IBG112039	3576	11	48,397	48,943	33	546	1	OR4A47	-0.59	-0.43	-1.76
IBG1448951	4442	14	19,848	20,420	17	573	3	10 genes (including several OR genes)[a]	-0.97	-1.05	-2.83
IBG143577	4010	2	132,731	133,354	120	622	3	ANKRD30BL,GPR39,MIR663B	-1.73	-1.54	-2.09
IBG112079	3906	8	105,737	106,407	147	670	3	ZFPM2	-1.56	-1.65	-1.98
IBG111829	2856	11	49,596	50,283	56	687	3	LOC440040,LOC441601,OR4C12,OR4C13	-1.93	-1.74	-2.78
IBG112389	4048	5	45,672	46,399	35	727	3	HCN1	-1.47	-1.57	-3.18
IBG145160	4499	11	54,794	56,004	190	1,209	3	30 genes (including several OR and TRIM genes)[b]	-2.21	-2.05	-3.63
IBG1451651	4499	11	54,794	56,004	190	1,209	3	30 genes (including several OR and TRIM genes)[b]	-1.83	-1.74	-2.14
IBG111948	3523	16	14,975	16,303	419	1,328	3	27 genes (including several microRNAs)[c]	-1.55	-1.53	-1.61

Table 2. Large annotated CNV events (>500 kb) detected in RD cases. When a CNV is annotated to more than five RefSeq genes, these are reported in a footnote (see below). All the CNVs partially overlapped or encompassed the genes to which they were annotated. All the positions are expressed in hg 19 coordinates.

[a] BMS1P17, BMS1P18, OR11H2, OR4K1, OR4K2, OR4K5, OR4M1, OR4N2, OR4Q3, POTEM.

[b] OR10AG1, OR4A15, OR4A16, OR4C11, OR4C15, OR4C16, OR4C6, OR4P4, OR4S2, OR5AS1, OR5D13, OR5D14, OR5D16, OR5D18, OR5F1, OR5I1, OR5J2, OR5L1, OR5L2, OR5T2, OR5W2, OR7E5P, OR8H2, OR8H3, OR8I2, OR8J3, OR8K5, TRIM48, TRIM51, TRIM51HP.

[c] ABCC1, ABCC6, C16orf45, FOPNL, KIAA0430, LOC100288162, MIR3179-1, MIR3179-2, MIR3179-3, MIR3180-1, MIR3180-2, MIR3180-3, MIR3180-4, MIR484, MIR6506, MIR6511A-2, MIR6511B-1, MIR6770-2, MPV17L, MYH11, NDE1, NOMO1, NPIPA1, NPIPA5, NTAN1, PDXDC1, RRN3.

Subject	Family	Chr	Start (kb)	End (kb)	SNPs	Length (kb)	CopyN	Confidence	Gene	PC1	IQadjPC1	IBGdiscr
IBG1123751	4034	18	22,928	22,956	7	28	3	19	ZNF521	-1.71	-1.55	-2.49
IBG1123761	4034	18	22,928	23,023	22	95	3	68	ZNF521	-1.57	-0.89	-1.45
IBG113128	4503	13	95,937	96,033	57	96	3	121	ABCC4	-1.8	-1.45	-2.63
IBG113129	4503	13	95,937	96,033	57	96	3	150	ABCC4	-1.55	-2.01	-1.5
IBG145160	4499	11	54,794	56,004	190	1,209	3	391	30 genes (including several OR and TRIM genes)[a]	-2.21	-2.05	-3.63
IBG1451651	4499	11	54,794	56,004	190	1,209	3	551	30 genes (including several OR and TRIM genes)[a]	-1.83	-1.74	-2.14
IBG142799	3514	6	145,148	145,175	15	27	3	30	UTRN	-1.62	-1.35	-2.12
IBG142797	3514	6	145,148	145,175	15	27	3	29	UTRN	-1.84	-2	-1.66
IBG142799	3514	1	225,391	225,454	14	63	3	37	DNAH14	-1.62	-1.35	-2.12
IBG142797	3514	1	225,391	225,454	14	63	3	39	DNAH14	-1.84	-2	-1.66
IBG111728	2615	4	165,577	165,587	8	10	1	32	MIR5684	-1.97	-1.56	-2.3
IBG111729	2615	4	165,577	165,587	8	10	1	23	MIR5684	-1.56	-1.37	-1.54

Table 3. Annotated CNVs shared between two or more affected co-siblings in ten families presenting more than one RD case. Some of these calls were detected also in non-affected participants, which are not shown in the present table (see *CNVs shared between RD cases* paragraph for further details). When a CNV is annotated to more than five RefSeq genes, these are reported in a footnote (see below). All the CNVs partially overlapped or encompassed the genes to which they were annotated. All the positions are expressed in hg 19 coordinates.

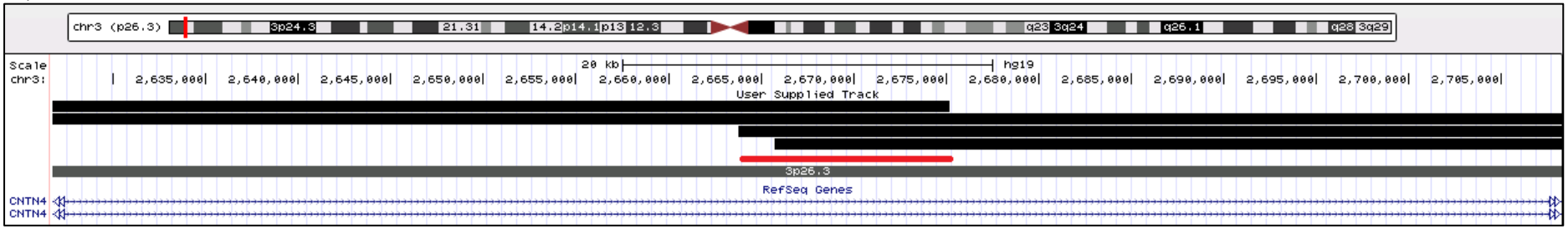
[a] OR10AG1, OR4A15, OR4A16, OR4C11, OR4C15, OR4C16, OR4C6, OR4P4, OR4S2, OR5AS1, OR5D13, OR5D14, OR5D16, OR5D18, OR5F1, OR5I1, OR5J2, OR5L1, OR5L2, OR5T2, OR5W2, OR7E5P, OR8H2, OR8H3, OR8I2, OR8J3, OR8K5, TRIM48, TRIM51, TRIM51HP.

Chr	Start (bp)	Stop (bp)	Kb	SNPs	p-val (PC1)	p-val (IQadjPC1)	Effect[a]	Freq (%)[b]	Gene[c]
3	2,663,757	2,675,189	11	13	[0.1; 0.12]	[3.0; 6.1] x 10^{-3}	+	0.6	CNTN4
5	36,449,552	36,461,331	12	12	[2.5; 4.3] x 10^{-3}	[2.5; 5.2] x 10^{-3}	+	0.4	
6	168,579,302	168,595,832	16	12	[5.8; 9.8] x 10^{-3}	[0.7; 2.1] x 10^{-3}	-	3.0	
10	68,223,696	68,242,672	19	9	[1.6; 1.9] x 10^{-2}	[4.0; 4.7] x 10^{-3}	-	0.4	CTNNA3
11	55,241,556	55,362,955	121	28	[0.5; 1.0] x 10^{-2}	[2.0; 4.8] x 10^{-3}	-	0.4	OR4C15, OR4C16

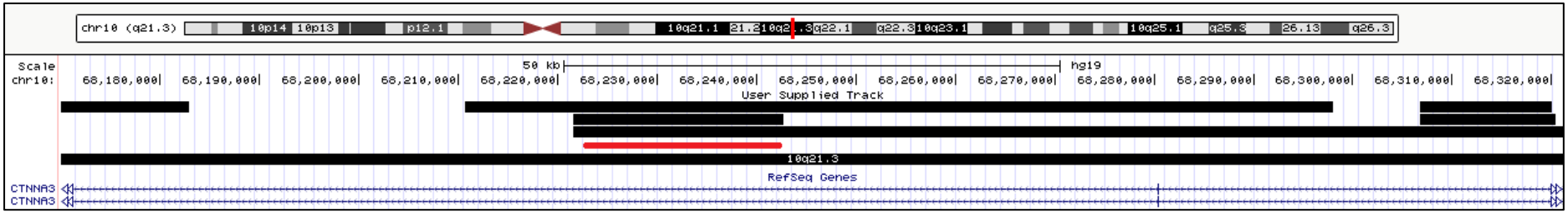
Table 4. Regions of CNV overlap showing the most significant associations with PC1/IQ-adjusted PC1 in the GWAS meta-analysis with CNV state (PLINK QFAM). All the regions of overlap of two or more CNVs (with at least one CNV call in CLDRC-RD and one in CLDRC-ADHD), showing at least two consecutive SNPs with association $p < 0.005$ with PC1 or IQ-adjusted PC1, are reported. The results of this meta-analysis on an individual probe basis are reported in detail in Table S3 a, b. All the positions are expressed in hg 19 coordinates.

[a] Effect of the CNV+ state, irrespective of the copy number, on PC1 and IQ-adjusted PC1. [b] Frequency (%) of the CNV+ state in the CLDRC dataset. [c] Genes overlapped/encompassed by the region reported.

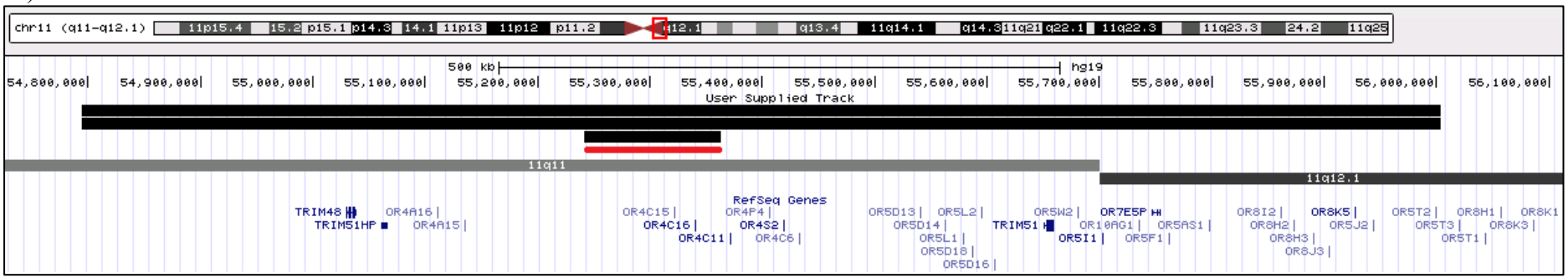
2a)


chr3 (p26.3)
Scale
chr3:
20 kb
hg19
User Supplied Track
3p26.3
RefSeq Genes
CNTN4
CNTN4

2b)


chr10 (q21.3)
Scale
chr10:
50 kb
hg19
User Supplied Track
10q21.3
RefSeq Genes
CTNNA3
CTNNA3

2c)


chr11 (q11-q12.1)
500 kb
hg19
User Supplied Track
11q11
11q12.1
RefSeq Genes
TRIM48
TRIM51HP
OR4A16
OR4A15
OR4C15
OR4P4
OR4C16
OR4S2
OR4C11
OR4C6
OR5D13
OR5L2
OR5W2
OR7E5P
OR5D14
TRIM51
OR10AG1
OR5AS1
OR5L1
OR5I1
OR5F1
OR5D18
OR5D16
OR8I2
OR8K5
OR5T2
OR8H1
OR8K1
OR8H2
OR5J2
OR5T3
OR8K3
OR8H3
OR5T1
OR8J3

2d)

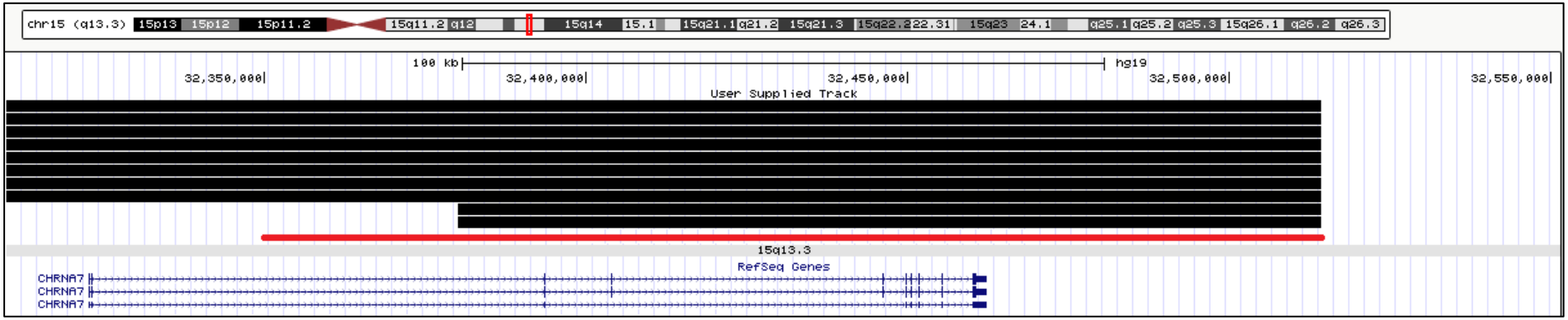


Figure 2. Regions of CNV overlap (labelled in red) associated with PC1/IQ-adjusted PC1 in the GWAS with CNV state (PLINK QFAM analysis). **a)** chr3:2,663,757-2,675,189, lying within *CNTN4* (3p26.3); **b)** chr10:68,223,696-68,242,672 within *CTNNA3* (10q21.3); **c)** chr11:55,241,556-55,362,955, encompassing *OR4C15* and *OR4C16* (11q11); **d)** chr15:32,350,775-32,514,341, partially overlapping *CHRNA7* (15q13.3). Figures **a, b, c** show the strongest associated regions overlapping with genes in this analysis, while Figure **d** illustrates the only associated region overlapping with known neuropsychiatric CNVs. Individual CNV calls are represented by black horizontal lines.

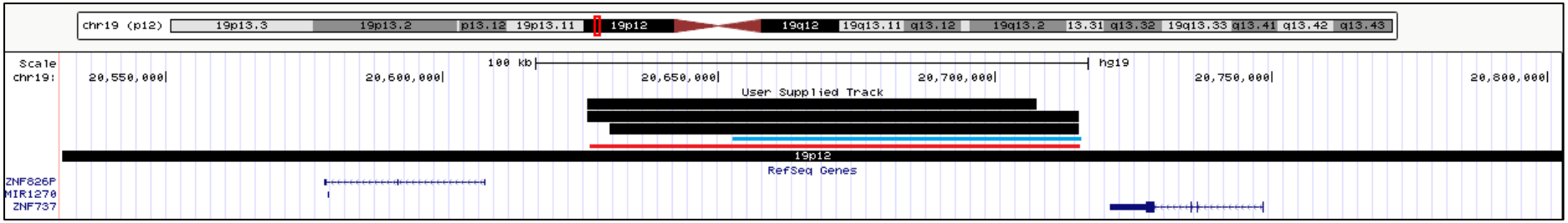


Figure 3. 19p12 region associated with PC1/IQ-adjusted PC1 in the GWAS with probe intensity data (FamCNV analysis, labelled in light blue) and overlapping region associated in the GWAS with CNV state (PLINK QFAM analysis, labelled in red). Black horizontal lines represent the three types of deletions detected in this region (reported in Table S2f).

Discussion

In the context of CNV research on reading and language, the present work presents three points of novelty which are worth stressing.

For the first time, we investigated the effect of CNVs on continuous reading and language traits, with enrichment for the lower tail of their distribution. This approach is more suited to the genetic background of a complex trait, compared to the assessment of co-segregation between CNVs and disorders in single families. A similar scope was conceived in a recent study by Stefansson and colleagues (2014), which investigated the effect of candidate neuropsychiatric CNVs on different cognitive traits in a big sample of the Icelandic population. However, this study analyzed a broader spectrum of general cognitive abilities in a wider range of variation (including general population controls) and was not aimed at assessing in detail a wealth of reading and language domains, as we did in our study.

Second, we derived and analyzed a principal component score, representing a substantial proportion of the shared variance in reading and language skills, both dependent (PC1) and independent of general cognitive abilities (IQ-adjusted PC1).

Third, to detect effects of CNVs on continuous reading and language traits, we used two complementary approaches: one aimed at detecting copy-number dependent effects and one aimed at detecting associations with the CNV state at each probe, irrespective of the abnormal copy number. These analyses were run in order to identify potential CNVs with reduced penetrance and variable expressivity on our traits of interest, and were in turn complementary to our classical analysis of co-segregation between CNVs and RD status in each sibship, aimed at detecting variants with full penetrance and expressivity.

In our dataset of subjects affected by RD/ADHD and their unaffected siblings, we did not identify any significant correlation between CNV genomic burden -both in terms of total length and of total number of CNVs per subject- and our PC scores representing reading-language performance. This is in line with a previous CNV study which detected no significant difference in the genomic burden of large rare CNVs between RD cases and controls (Girirajan et al., 2011). Nonetheless, our result is in partial contrast with a recent study reporting an increased CNV burden in SLI cases compared to controls (Simpson et al., 2015). This discrepancy may be explained through the hypothesis that CNVs contribute specifically to SLI but not to RD and that, similarly, they do not severely affect cognitive

domains shared between reading and language. Further detailed analyses in independent datasets will be needed to clarify this aspect.

In this study we detected a CNV which co-segregated with the dyslexic status in a family with two RD cases -including the most severely impaired subject in our dataset- and one unaffected sibling. This large CNV event spanned ~1.2 Mb in the pericentromeric region 11q11-q12.1, covering several *OR* (olfactory receptors) and *TRIM* (tripartite motif protein) genes. While TRIM proteins are not well characterized, the role of olfactory receptors in triggering odor perception signals in sensory neurons is well known. Due to their biological function, *OR* genes have been well conserved during animal evolution but selective pressures at these loci have relaxed in the human lineage (Pierron et al., 2013). Interestingly, olfactory bulbs dysgenesis/agenesis has been previously implicated in ASD (Brang & Ramachandran, 2010) and reduced volumes have been reported in schizophrenic patients (Turetsky et al., 2000). However, the partial overlap of this CNV with a centromeric region and the relaxed selection at the *OR* loci suggest caution in the biological interpretation of this variant.

Other CNVs shared between cases were detected, overlapping potential susceptibility genes. Among these, the most interesting candidates -shared by a pair of affected siblings in a family but not detected in any control- were two heterozygous duplications, one encompassing 9 exons in the 3' terminal region of the *UTRN* gene (utrophin, or dystrophin-related protein 1, 6q24.2) and the other one overlapping 12 exons within *DNAH14* (dynein axonemal heavy chain 14, 1q42.12). Utrophin is a large skeletal muscle protein contributing to post-synaptic membrane maintenance and to clustering of acetylcholine receptors in the neuromuscular synapses, and possibly playing a role in anchoring the cytoskeleton to the plasma membrane. It is also expressed in the CNS (Blake et al., 1995) and shows strong structural and functional similarities with the dystrophin protein, which is at the basis of Duchenne Muscular Dystrophy (DMD). Interestingly, DMD cases sometimes show cognitive impairment, reading and language deficits, in addition to typical muscular weakness and progressive paralysis (Perronnet & Vaillend, 2010). Dyneins are microtubule-associated motor proteins fundamental for several cellular processes, including cell motility, through cilia. Independent studies have reported evidence of involvement in cilia-related processes for two RD candidate genes: *DYX1C1* plays a role in cilia assembly, growth and function (Chandrasekar et al., 2013; Tarkar et al., 2013), while *DCDC2* affects primary cilia structure and signaling (Massinen et al., 2011). In view of this evidence, it has been hypothesized that dyslexia may sometimes be a form of ciliopathy (Chandrasekar et al., 2013), involving abnormal neuronal

development and migration (Massinen et al., 2011). A small (~10 kb) heterozygous deletion, overlapping *MIR5684* (microRNA 5684, 4q32.3), was also detected in two affected siblings but in no controls. This also constitutes a plausible candidate, as microRNAs have been implicated in the etiology of RD and SLI, through altered post-transcriptional regulation of several genes involved in the axonal guidance pathway (Rudov et al., 2013).

Pathway-based enrichment test of CNV calls detected in RD cases revealed no significant associations for three candidate gene sets representing mainstream hypotheses on the etiology of RD and SLI, namely axon guidance, neuron migration and steroids-related processes. This is in line with the pathway enrichment test based on SNP associations in our GWASMA (Chapter 3), and suggests that putative pathological CNVs alone do not heavily affect these pathways. Nonetheless, we cannot exclude that SNPs and CNVs may still contribute to alter these molecular networks jointly with other kind of variants, such as rare mutations and Short Tandem Repeats (STRs), exerting very small individual effects.

An interesting aspect of this study is the use of two complementary strategies for genome-wide association testing between CNVs and our principal component reading-language scores, PC1 and IQ-adjusted PC1. The first of these analyses, based on PennCNV and then PLINK QFAM, was aimed at detecting associations in regions of overlap of CNV calls, irrespective of the abnormal copy number state (be it 0, 1, 3 or 4). The second analysis, in FamCNV, assessed copy number- (or allele dosage-) dependent associations between DNA array intensity data and PC1/IQ-adjusted PC1. These are complementary and practical strategies to detect effects of CNVs on continuous traits: in recent studies a copy number-dependent effect was reported for continuous traits such as Body Mass Index (Falchi et al., 2014) and neural phenotypes, including corpus callosum volume and white matter volumes in the temporal lobe (Stefansson et al., 2014); conversely, deletions and reciprocal duplications in specific regions often result in similar clinical and phenotypic manifestations, as in the case of autism, language/developmental delays, and other psychiatric disorders (Burnside et al., 2011; Grayton et al., 2012; Griswold et al., 2012; Malhotra & Sebat, 2012; Weiss et al., 2008). Even if both analyses were run probe-by-probe, results were interpreted in terms of consecutive probes showing significant associations, which was more appropriate for the kind of variants investigated.

Although no associations survived correction for multiple testing in PLINK QFAM meta-analysis, some of the top associated regions lay within genes. A ~12 kb CNV overlap, associated with IQ-adjusted PC1, lay in an intronic region within *CNTN4* (contactin 4,

3p26.3; Figure 2a). This overlap was shared by three heterozygous duplications and one heterozygous deletion, which were all showing concordant positive effects on PC score. Contactins are Ig cell adhesion molecules with a fundamental role in neuronal development and plasticity. CNVs and structural rearrangements disrupting contactin 4 have been implicated in severe neurodevelopmental disorders such as ASD (Glessner et al., 2009; Guo et al., 2012; Roohi et al., 2009) and DD (Fernandez et al., 2004; Fernandez et al., 2008). Interestingly, the associated region detected in the present work overlaps with CNVs reported in ASD cases in two previous studies (Guo et al., 2012; Roohi et al., 2009), and contactin 4 is widely expressed in the brain, particularly in cerebellum, thalamus, amygdala, and cerebral cortex (Guo et al., 2012; Zuko et al., 2013). However, this association was weaker with PC1, falling short out of significance ($p \sim 0.1$).

Another intronic CNV overlap of ~19 kb, associated with both PC1 and IQ-adjusted PC1, was found within *CTNNA3* (catenin alpha 3, 10q21.3; Figure 2b). This region resulted from the overlap of three deletions and showed a negative effect on PC scores. The same trend was observed in FamCNV analysis of this region, although associations were not significant (data not shown). α-catenins have a crucial role in cell adhesion and *CTNNA3* has been implicated in ASD etiology both through CNV studies (Bacchelli et al., 2014; Levy et al., 2011; Nava et al., 2014) and through GWAS studies (Wang et al., 2009; Weiss et al., 2009). Interestingly, our associated region partially overlaps an inherited compound heterozygous deletion encompassing exon 11, found in an ASD patient (Bacchelli et al., 2014). Expression of *CTNNA3* in mouse hippocampus and cortex at postnatal day 0 suggests a specific neuronal role at very early developmental stages (Bacchelli et al., 2014). This makes this gene a very interesting candidate susceptibility locus for neurodevelopmental traits like reading and language.

Another associated region in PLINK QFAM meta-analysis which annotated to genes was found on 11q11, at the overlap between two big heterozygous duplications reported above and a heterozygous deletion encompassing genes *OR4C15* and *OR4C16* (olfactory receptors 15 and 16, family 4, subfamily C; Figure 2c). This region showed nominally significant association only with IQ-adjusted PC1 -with both duplications and deletion showing a negative effect on this score- and lay within a pericentromeric region, encompassing genes under relaxed selective pressure. Therefore, caution is suggested in the interpretation of this result, as discussed above. More in general, low frequencies (<1%) of CNV+ state in most of the top associated regions in this analysis (see Table 4) suggest prudence in the interpretation

of these associations, especially for those CNV overlaps between participants with extreme PC1/IQ-adjusted PC1 scores (as in the case of 11q11 duplications, see Table 2).

In the context of QFAM analysis, we also assessed CNV overlaps in regions previously reported to be disrupted by CNVs in RD, SLI or more severe neuropsychiatric disorders (see *Results* section). Among these, a ~164 kb region of overlap between nine heterozygous duplications and one heterozygous deletion, encompassing several exons in the 3' region of *CHRNA7* (cholinergic nicotinic receptor alpha 7, 15q13.3; Figure 2d), presented nominally significant association with PC1 in the CLDRC-RD subset (while no CNV calls were detected in CLDRC-ADHD). The association only approached significance after IQ-adjustment ($p \sim 0.055$-0.1) and the CNV state exerted a positive effect on PC1/IQ-adjusted PC1, with both deletion and duplications showing the same trend. This CNV looks a very convincing candidate for affecting reading and language skills for a number of reasons. Primarily, nicotinic cholinergic receptors are ligand-gated ion channels that mediate fast signal transmission at synapses and are ubiquitously expressed in the CNS (Helbig et al., 2009). Secondarily, the 15q13.3 region is a hotspot of neuropsychiatric CNVs, which have been implicated in several disorders including schizophrenia (Malhotra & Sebat, 2012; Stefansson et al., 2008), ASD (Grayton et al., 2012; Griswold et al., 2012; Malhotra & Sebat, 2012), ADHD (Williams et al., 2012) and epilepsy (Helbig et al., 2009). Moreover, a CNV encompassing *CHRNA7* was suggested to contribute to the disruption of the synaptic pathway in a patient with ID and SLI (Chilian et al., 2013). A recent CNV study also reported *CHRNA7* among the genes overlapped by CNVs in a group of unrelated SLI cases, and a significant overrepresentation of the GO category *acetylcholine binding* in a pathway-based analysis of these CNVs (Simpson et al. 2015). CNVs encompassing this gene were also tested for effects on general cognitive abilities, including school history of mathematical and reading difficulties, in a big Icelandic population sample, but no associations were reported (Stefansson et al., 2014).

Similarly to PLINK QFAM analysis, FamCNV meta-analysis did not reveal any genome-wide significant association surviving correction for multiple testing. However, we found a series of eight contiguous SNPs associated with both PC1 and IQ-adjusted PC1 in the CLDRC-RD analysis, ~6 kb downstream of *ZNF737* (zinc finger protein 737, 19p12, Figure 3). This ~58 kb region lay within a ~80 kb deletion very frequent in our dataset and the association was also observed at the nominal significance level in the PLINK QFAM analysis of CLDRC-RD. Both FamCNV and QFAM analysis indicated a positive effect of this

deletion on PC1/IQ-adjusted PC1. Zinc finger protein 737 has not been functionally characterized, but the presence of a zinc finger domain suggests a possible involvement in transcriptional regulation. Interestingly, a microdeletion within another zinc finger gene, *ZNF277*, was recently suggested as susceptibility CNV for SLI (Ceroni et al., 2014).

In spite of the interesting suggestive associations discussed above, the modest sample size, the absence of a replication sample and of a molecular validation of CNV breakpoints constitute limitations for the present study. Therefore, further analyses in larger datasets, including the localization of CNV breakpoints, are warranted to validate and extend such associations. Also, RD cases definition was somehow arbitrary. Nonetheless, for completeness of our analysis, we decided to use it to assess co-segregation with CNVs in the sibships. As there is no universal agreement on the diagnostic definition of dyslexia in the scientific community (Peterson and Pennington, 2012; Raskind et al., 2013) we decided to use a "performance only"-based criterion, classifying all the participants in the lowest 10% of the IBG discriminant score distribution as RD cases, and considering them as representative of reading impairment. In spite of these limitations, the methodological approach used in this study represents an interesting strategy to investigate the effects of CNVs on neuropsychiatric traits. We believe that this comprehensive approach should be used in future CNV research in the genetics of language and neuropsychiatric traits in general, possibly applying it to datasets larger than the one used in this work. This may allow to identify new structural variants with subtle effects and to possibly clarify part of the missing heritability in reading and language.

Supplementary Material

- *S1: Supplementary Methods.* Details on IBG discriminant score; pairwise phenotypic correlations of PC1, IQadjPC1 and IBGdiscr; Principal Component Analysis of LRR intensity data from DNA array.
- *S2: CNV calls of interest.* CNV calls within genes implicated in reading and language; CNV calls overlapping genes previously detected in CNV studies of RD/SLI; CNV calls overlapping common neuropsychiatric CNVs; CNV calls overlapping/encompassing *CHRNA7* and *ZNF737*.
- *S3: Supplementary Results.* Results of FamCNV and PLINK QFAM GWAS meta-analyses on an individual probe basis; details of associations in the *ZNF737* region; results of pathway-based analysis.

Acknowledgements

We would like to thank Dr Kai Wang for technical assistance using PennCNV. This work was supported by the Max Planck Society and the US National Institutes of Health (Grant ref: P50 HD027802). The authors declare no competing interests.

References

Anthoni, H., Zucchelli, M., Matsson, H., Muller-Myhsok, B., Fransson, I., Schumacher, J., Massinen, S., Onkamo, P., Warnke, A., Griesemann, H., Hoffmann, P., Nopola-Hemmi, J., Lyytinen, H., Schulte-Korne, G., Kere, J., Nothen, M.M. & Peyrard-Janvid, M. (2007) A locus on 2p12 containing the co-regulated MRPL19 and C2ORF3 genes is associated to dyslexia. *Hum Mol Genet,* **16,** 667-677.

Bacchelli, E., Ceroni, F., Pinto, D., Lomartire, S., Giannandrea, M., D'Adamo, P., Bonora, E., Parchi, P., Tancredi, R., Battaglia, A. & Maestrini, E. (2014) A CTNNA3 compound heterozygous deletion implicates a role for alphaT-catenin in susceptibility to autism spectrum disorder. *J Neurodev Disord,* **6,** 17.

Barry, J.G., Yasin, I. & Bishop, D.V. (2007) Heritable risk factors associated with language impairments. *Genes Brain Behav,* **6,** 66-76.

Bates, T.C., Luciano, M., Medland, S.E., Montgomery, G.W., Wright, M.J. & Martin, N.G. (2011) Genetic variance in a component of the language acquisition device: ROBO1 polymorphisms associated with phonological buffer deficits. *Behav Genet,* **41,** 50-57.

Blake, D.J., Schofield, J.N., Zuellig, R.A., Górecki, D.C., Phelps, S.R., Barnard, E.A., Edwards, Y.H. & Davies, K.E. (1995) G-utrophin, the autosomal homologue of dystrophin Dp116, is expressed in sensory ganglia and brain. *Proc Natl Acad Sci USA,* **92,** 3697-3701.

Brang, D. & Ramachandran, V.S. (2010) Olfactory bulb dysgenesis, mirror neuron system dysfunction, and autonomic dysregulation as the neural basis for autism. *Med Hypotheses,* **74,** 919-921.

Burnside, R., Pasion, R., Mikhail, F., Carroll, A., Robin, N., Youngs, E., Gadi, I., Keitges, E., Jaswaney, V., Papenhausen, P., Potluri, V., Risheg, H., Rush, B., Smith, J., Schwartz, S., Tepperberg, J. & Butler, M. (2011) Microdeletion/microduplication of proximal 15q11.2 between BP1 and BP2: a susceptibility region for neurological dysfunction including developmental and language delay. *Hum Genet,* **130,** 517-528.

Carrion-Castillo, A., Franke, B. & Fisher, S.E. (2013) Molecular Genetics of Dyslexia: An Overview. *Dyslexia,* **19,** 214-240.

Ceroni, F., Simpson, N.H., Francks, C., Baird, G., Conti-Ramsden, G., Clark, A., Bolton, P.F., Hennessy, E.R., Donnelly, P., Bentley, D.R., Martin, H., Parr, J., Pagnamenta, A.T., Maestrini, E., Bacchelli, E., Fisher, S.E. & Newbury, D.F. (2014) Homozygous microdeletion of exon 5 in ZNF277 in a girl with specific language impairment. *Eur J Hum Genet*, **22**, 1165–1171.

Chandrasekar, G., Vesterlund, L., Hultenby, K., Tapia-Paez, I. & Kere, J. (2013) The zebrafish orthologue of the dyslexia candidate gene DYX1C1 is essential for cilia growth and function. *PLoS ONE,* **8,** e63123.

Chilian, B., Abdollahpour, H., Bierhals, T., Haltrich, I., Fekete, G., Nagel, I., Rosenberger, G. & Kutsche, K. (2013) Dysfunction of SHANK2 and CHRNA7 in a patient with intellectual disability and language impairment supports genetic epistasis of the two loci. *Clin Genet,* **84,** 560-565.

Compton, D.L., DeFries, J.C. & Olson, R.K. (2001) Are RAN- and phonological awareness-deficits additive in children with reading disabilities? *Dyslexia,* **7,** 125-149.

Cope, N., Harold, D., Hill, G., Moskvina, V., Stevenson, J., Holmans, P., Owen, M.J., O'Donovan, M.C. & Williams, J. (2005) Strong Evidence That KIAA0319 on Chromosome 6p Is a Susceptibility Gene for Developmental Dyslexia. *Am J Hum Genet,* **76,** 581-591.

DeFries, J.C. (1985) Colorado Reading Project. Parkton, MD: York Press, pp. 107-122.

DeFries, J.C., Filipek, P.A., Fulker, D.W., Olson, R.K., Pennington, B.F., Smith, S.D. & Wise, B.W. (1997) Colorado Learning Disabilities Research Center. *Learn Disabil Multidisciplinary J,* **8,** 7-19.

Dunn, L.M. & Markwardt, F.C. (1970) *Examiner's Manual: Peabody Individual Achievement Test.,* American Guidance Service: Circle Pines, MN.

Eicher, J.D., Powers, N.R., Miller, L.L., Akshoomoff, N., Amaral, D.G., Bloss, C.S., Libiger, O., Schork, N.J., Darst, B.F., Casey, B.J., Chang, L., Ernst, T., Frazier, J., Kaufmann, W.E., Keating, B., Kenet, T., Kennedy, D., Mostofsky, S., Murray, S.S., Sowell, E.R., Bartsch, H., Kuperman, J.M., Brown, T.T., Hagler, D.J., Jr., Dale, A.M., Jernigan, T.L., Pourcain, B.S., Smith, G.D., Ring, S.M. & Gruen, J.R. (2013) Genome-Wide Association Study of Shared Components of Reading Disability and Language Impairment. *Genes Brain Behav*, **12**, 792-801.

Eleftherohorinou, H., Andersson-Assarsson, J.C., Walters, R.G., El-Sayed Moustafa, J.S., Coin, L., Jacobson, P., Carlsson, L.M., Blakemore, A.I., Froguel, P., Walley, A.J. & Falchi, M. (2011) famCNV: copy number variant association for quantitative traits in families. *Bioinformatics,* **27,** 1873-1875.

Elia, J., Glessner, J.T., Wang, K., Takahashi, N., Shtir, C.J., Hadley, D., Sleiman, P.M.A., Zhang, H., Kim, C.E., Robison, R., Lyon, G.J., Flory, J.H., Bradfield, J.P., Imielinski, M., Hou, C., Frackelton, E.C., Chiavacci, R.M., Sakurai, T., Rabin, C., Middleton, F.A., Thomas, K.A., Garris, M., Mentch, F., Freitag, C.M., Steinhausen, H.-C., Todorov, A.A., Reif, A., Rothenberger, A., Franke, B., Mick, E.O., Roeyers, H., Buitelaar, J., Lesch, K.-P., Banaschewski, T., Ebstein, R.P., Mulas, F., Oades, R.D., Sergeant, J., Sonuga-Barke, E., Renner, T.J., Romanos, M., Romanos, J., Warnke, A., Walitza, S., Meyer, J., Palmason, H., Seitz, C., Loo, S.K., Smalley, S.L., Biederman, J., Kent, L., Asherson, P., Anney, R.J.L., Gaynor, J.W., Shaw, P., Devoto, M., White, P.S., Grant, S.F.A., Buxbaum, J.D., Rapoport, J.L., Williams, N.M., Nelson, S.F., Faraone, S.V. & Hakonarson, H. (2012) Genome-wide copy number variation study associates metabotropic glutamate receptor gene networks with attention deficit hyperactivity disorder. *Nat Genet*, **44**, 78-84.

Falchi, M., El-Sayed Moustafa, J.S., Takousis, P., Pesce, F., Bonnefond, A., Andersson-Assarsson, J.C., Sudmant, P.H., Dorajoo, R., Al-Shafai, M.N., Bottolo, L., Ozdemir, E., So, H.-C., Davies, R.W., Patrice, A., Dent, R., Mangino, M., Hysi, P.G., Dechaume, A., Huyvaert, M., Skinner, J., Pigeyre, M., Caiazzo, R., Raverdy, V., Vaillant, E., Field, S., Balkau, B., Marre, M., Visvikis-Siest, S., Weill, J., Poulain-Godefroy, O., Jacobson, P., Sjostrom, L., Hammond, C.J., Deloukas, P., Sham, P.C., McPherson, R., Lee, J., Tai, E.S., Sladek, R., Carlsson, L.M.S., Walley, A., Eichler, E.E., Pattou, F., Spector, T.D. & Froguel, P. (2014) Low copy number of the salivary amylase gene predisposes to obesity. *Nat Genet,* **46,** 492-497.

Fernandez, T., Morgan, T., Davis, N., Klin, A., Morris, A., Farhi, A., Lifton, R.P. & State, M.W. (2004) Disruption of Contactin 4 (CNTN4) Results in Developmental Delay and Other Features of 3p Deletion Syndrome. *Am J Hum Genet,* **74,** 1286-1293.

Fernandez, T., Morgan, T., Davis, N., Klin, A., Morris, A., Farhi, A., Lifton, R.P. & State, M.W. (2008) Disruption of Contactin 4 (CNTN4) Results in Developmental Delay and Other Features of 3p Deletion Syndrome. *Am J Hum Genet,* **82,** 1385.

Fisher, S.E. & DeFries, J.C. (2002) Developmental dyslexia: genetic dissection of a complex cognitive trait. *Nat Rev Neurosci,* **3,** 767-780.

Fisher, S.E. & Scharff, C. (2009) FOXP2 as a molecular window into speech and language. *Trends Genet,* **25,** 166-177.

Francks, C., Paracchini, S., Smith, S.D., Richardson, A.J., Scerri, T.S., Cardon, L.R., Marlow, A.J., MacPhie, I.L., Walter, J., Pennington, B.F., Fisher, S.E., Olson, R.K., DeFries, J.C., Stein, J.F. & Monaco, A.P. (2004) A 77-Kilobase Region of Chromosome 6p22.2 Is Associated with Dyslexia in Families From the United Kingdom and From the United States. *Am J Hum Genet,* **75,** 1046-1058.

Friend, A. & Olson, R.K. (2010) Phonological Spelling and Reading Deficits in Children with Spelling Disabilities. *Sci Stud Read,* **12,** 90-105.

Gialluisi, A., Newbury, D.F., Wilcutt, E.G., Olson, R.K., DeFries, J.C., Brandler, W.M., Pennington, B.F., Smith, S.D., Scerri, T.S., Simpson, N.H., The, S.L.I.C., Luciano, M., Evans, D.M., Bates, T.C., Stein, J.F., Talcott, J.B., Monaco, A.P., Paracchini, S., Francks, C. & Fisher, S.E. (2014) Genome-wide screening for DNA variants associated with reading and language traits. *Genes Brain Behav***, 13**, 686-701.

Girirajan, S., Brkanac, Z., Coe, B.P., Baker, C., Vives, L., Vu, T.H., Shafer, N., Bernier, R., Ferrero, G.B., Silengo, M., Warren, S.T., Moreno, C.S., Fichera, M., Romano, C., Raskind, W.H. & Eichler, E.E. (2011) Relative burden of large CNVs on a range of neurodevelopmental phenotypes. *PLoS Genet,* **7,** e1002334.

Glessner, J.T., Wang, K., Cai, G., Korvatska, O., Kim, C.E., Wood, S., Zhang, H., Estes, A., Brune, C.W., Bradfield, J.P., Imielinski, M., Frackelton, E.C., Reichert, J., Crawford, E.L., Munson, J., Sleiman, P.M.A., Chiavacci, R., Annaiah, K., Thomas, K., Hou, C., Glaberson, W., Flory, J., Otieno, F., Garris, M., Soorya, L., Klei, L., Piven, J., Meyer, K.J., Anagnostou, E., Sakurai, T., Game, R.M., Rudd, D.S., Zurawiecki, D., McDougle, C.J., Davis, L.K., Miller, J., Posey, D.J., Michaels, S., Kolevzon, A.,

Silverman, J.M., Bernier, R., Levy, S.E., Schultz, R.T., Dawson, G., Owley, T., McMahon, W.M., Wassink, T.H., Sweeney, J.A., Nurnberger, J.I., Coon, H., Sutcliffe, J.S., Minshew, N.J., Grant, S.F.A., Bucan, M., Cook, E.H., Buxbaum, J.D., Devlin, B., Schellenberg, G.D. & Hakonarson, H. (2009) Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature,* **459,** 569-573.

Good, C.D., Johnsrude, I., Ashburner, J., Henson, R.N., Friston, K.J. & Frackowiak, R.S. (2001) Cerebral asymmetry and the effects of sex and handedness on brain structure: a voxel-based morphometric analysis of 465 normal adult human brains. *Neuroimage,* **14,** 685-700.

Grayton, H.M., Fernandes, C., Rujescu, D. & Collier, D.A. (2012) Copy number variations in neurodevelopmental disorders. *Prog Neurobiol,* **99,** 81-91.

Griswold, A.J., Ma, D., Cukier, H.N., Nations, L.D., Schmidt, M.A., Chung, R.-H., Jaworski, J.M., Salyakina, D., Konidari, I., Whitehead, P.L., Wright, H.H., Abramson, R.K., Williams, S.M., Menon, R., Martin, E.R., Haines, J.L., Gilbert, J.R., Cuccaro, M.L. & Pericak-Vance, M.A. (2012) Evaluation of copy number variations reveals novel candidate genes in autism spectrum disorder-associated pathways. *Hum Mol Genet,* **21,** 3513-3523.

Guo, H., Xun, G., Peng, Y., Xiang, X., Xiong, Z., Zhang, L., He, Y., Xu, X., Liu, Y., Lu, L., Long, Z., Pan, Q., Hu, Z., Zhao, J. & Xia, K. (2012) Disruption of Contactin 4 in two subjects with autism in Chinese population. *Gene,* **505,** 201-205.

Hannula-Jouppi, K., Kaminen-Ahola, N., Taipale, M., Eklund, R., Nopola-Hemmi, J., Kääriäinen, H. & Kere, J. (2005) The axon guidance receptor gene ROBO1 is a candidate gene for developmental dyslexia. *PLoS genetics,* **1,** e50.

Harlaar, N., Hayiou-Thomas, M.E., Dale, P.S. & Plomin, R. (2008) Why Do Preschool Language Abilities Correlate With Later Reading? A Twin Study. *J Speech Lang Hear Res,* **51,** 688-705.

Helbig, I., Mefford, H.C., Sharp, A.J., Guipponi, M., Fichera, M., Franke, A., Muhle, H., de Kovel, C., Baker, C., von Spiczak, S., Kron, K.L., Steinich, I., Kleefuss-Lie, A.A., Leu, C., Gaus, V., Schmitz, B., Klein, K.M., Reif, P.S., Rosenow, F., Weber, Y., Lerche, H., Zimprich, F., Urak, L., Fuchs, K., Feucht, M., Genton, P., Thomas, P., Visscher, F., de Haan, G.J., Moller, R.S., Hjalgrim, H., Luciano, D., Wittig, M., Nothnagel, M., Elger, C.E., Nurnberg, P., Romano, C., Malafosse, A., Koeleman, B.P., Lindhout, D., Stephani, U., Schreiber, S., Eichler, E.E. & Sander, T. (2009) 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. *Nat Genet,* **41,** 160-162.

Lee, P.H., O'Dushlaine, C., Thomas, B. & Purcell, S.M. (2012) INRICH: interval-based enrichment analysis for genome-wide association studies. *Bioinformatics,* **28,** 1797-1799.

Levy, D., Ronemus, M., Yamrom, B., Lee, Y.-h., Leotta, A., Kendall, J., Marks, S., Lakshmi, B., Pai, D., Ye, K., Buja, A., Krieger, A., Yoon, S., Troge, J., Rodgers, L., Iossifov, I. & Wigler, M. (2011) Rare De Novo and Transmitted Copy-Number Variation in Autistic Spectrum Disorders. *Neuron,* **70,** 886-897.

Logan, J., Petrill, S., Flax, J., Justice, L., Hou, L., Bassett, A., Tallal, P., Brzustowicz, L. & Bartlett, C. (2011) Genetic Covariation Underlying Reading, Language and Related Measures in a Sample Selected for Specific Language Impairment. *Behav Genet,* **41,** 651-659.

Lombardo, M.V., Ashwin, E., Auyeung, B., Chakrabarti, B., Taylor, K., Hackett, G., Bullmore, E.T. & Baron-Cohen, S. (2012) Fetal testosterone influences sexually dimorphic gray matter in the human brain. *J Neurosci,* **32,** 674-680.

Luciano, M., Evans, D.M., Hansell, N.K., Medland, S.E., Montgomery, G.W., Martin, N.G., Wright, M.J. & Bates, T.C. (2013) A genome-wide association study for reading and language abilities in two population cohorts. *Genes Brain Behav,* **12,** 645-652.

Malhotra, D. & Sebat, J. (2012) CNVs: Harbingers of a Rare Variant Revolution in Psychiatric Genetics. *Cell,* **148,** 1223-1241.

Massinen, S., Hokkanen, M.E., Matsson, H., Tammimies, K., Tapia-Paez, I., Dahlstrom-Heuser, V., Kuja-Panula, J., Burghoorn, J., Jeppsson, K.E., Swoboda, P., Peyrard-Janvid, M., Toftgard, R., Castren, E. & Kere, J. (2011) Increased expression of the dyslexia candidate gene DCDC2 affects length and signaling of primary cilia in neurons. *PLoS ONE,* **6,** e20580.

Massinen, S., Tammimies, K., Tapia-Paez, I., Matsson, H., Hokkanen, M.E., Soderberg, O., Landegren, U., Castren, E., Gustafsson, J.A., Treuter, E. & Kere, J. (2009) Functional interaction of DYX1C1 with

estrogen receptors suggests involvement of hormonal pathways in dyslexia. *Hum Mol Genet,* **18,** 2802-2812.

McArthur, G.M., Hogben, J.H., Edwards, V.T., Heath, S.M. & Mengler, E.D. (2000) On the "specifics" of specific reading disability and specific language impairment. *J Child Psychol Psychiatry,* **41,** 869–874.

Meng, H., Smith, S.D., Hager, K., Held, M., Liu, J., Olson, R.K., Pennington, B.F., DeFries, J.C., Gelernter, J., O'Reilly-Pol, T., Somlo, S., Skudlarski, P., Shaywitz, S.E., Shaywitz, B.A., Marchione, K., Wang, Y., Paramasivam, M., LoTurco, J.J., Page, G.P. & Gruen, J.R. (2005) DCDC2 is associated with reading disability and modulates neuronal development in the brain. *Proc Natl Acad Sci USA,* **102,** 17053-17058.

Nava, C., Keren, B., Mignot, C., Rastetter, A., Chantot-Bastaraud, S., Faudet, A., Fonteneau, E., Amiet, C., Laurent, C., Jacquette, A., Whalen, S., Afenjar, A., Perisse, D., Doummar, D., Dorison, N., Leboyer, M., Siffroi, J.-P., Cohen, D., Brice, A., Heron, D. & Depienne, C. (2014) Prospective diagnostic analysis of copy number variants using SNP microarrays in individuals with autism spectrum disorders. *Eur J Hum Genet,* **22,** 71-78.

Newbury, D.F. & Monaco, A.P. (2010) Genetic advances in the study of speech and language disorders. *Neuron,* **68,** 309-320.

Newbury, D.F., Paracchini, S., Scerri, T.S., Winchester, L., Addis, L., Richardson, A.J., Walter, J., Stein, J.F., Talcott, J.B. & Monaco, A.P. (2011) Investigation of dyslexia and SLI risk variants in reading- and language-impaired subjects. *Behav Genet,* **41,** 90-104.

Newbury, D.F., Winchester, L., Addis, L., Paracchini, S., Buckingham, L.L., Clark, A., Cohen, W., Cowie, H., Dworzynski, K., Everitt, A., Goodyer, I.M., Hennessy, E., Kindley, A.D., Miller, L.L., Nasir, J., O'Hare, A., Shaw, D., Simkin, Z., Simonoff, E., Slonims, V., Watson, J., Ragoussis, J., Fisher, S.E., Seckl, J.R., Helms, P.J., Bolton, P.F., Pickles, A., Conti-Ramsden, G., Baird, G., Bishop, D.V. & Monaco, A.P. (2009) CMIP and ATP2C2 modulate phonological short-term memory in language impairment. *Am J Hum Genet,* **85,** 264-272.

Nopola-Hemmi, J., Taipale, M., Haltia, T., Lehesjoki, A.-E., Voutilainen, A. & Kere, J. (2000) Two translocations of chromosome 15q associated with dyslexia. *J Med Genet,* **37,** 771-775.

Oksenberg, N. & Ahituv, N. (2013) The role of AUTS2 in neurodevelopment and human evolution. *Trends Genet,* **29,** 600-608.

Pagnamenta, A.T., Bacchelli, E., de Jonge, M.V., Mirza, G., Scerri, T.S., Minopoli, F., Chiocchetti, A., Ludwig, K.U., Hoffmann, P., Paracchini, S., Lowy, E., Harold, D.H., Chapman, J.A., Klauck, S.M., Poustka, F., Houben, R.H., Staal, W.G., Ophoff, R.A., O'Donovan, M.C., Williams, J., Nothen, M.M., Schulte-Korne, G., Deloukas, P., Ragoussis, J., Bailey, A.J., Maestrini, E. & Monaco, A.P. (2010) Characterization of a family with rare deletions in CNTNAP5 and DOCK4 suggests novel risk loci for autism and dyslexia. *Biol Psychiatry,* **68,** 320-328.

Paracchini, S., Steer, C.D., Buckingham, L.L., Morris, A.P., Ring, S., Scerri, T., Stein, J., Pembrey, M.E., Ragoussis, J., Golding, J. & Monaco, A.P. (2008) Association of the KIAA0319 Dyslexia Susceptibility Gene With Reading Skills in the General Population. *Am J Psychiatry,* **165,** 1576-1584.

Paracchini, S. (2011) Dissection of genetic associations with language-related traits in population-based cohorts. *J Neurodev Disord,* **3,** 365-373.

Pennington, B.F. (2006) From single to multiple deficit models of developmental disorders. *Cognition,* **101,** 385-413.

Pennington, B.F. & Bishop, D.V. (2009) Relations among speech, language, and reading disorders. *Annu Rev Psych,* **60,** 283-306.

Perronnet, C. & Vaillend, C. (2010) Dystrophins, Utrophins, and Associated Scaffolding Complexes: Role in Mammalian Brain and Implications for Therapeutic Strategies. *J Biomed Biotechnol,* **2010,** 19.

Peter, B., Matsushita, M., Oda, K. & Raskind, W. (2014) De novo microdeletion of BCL11A is associated with severe speech sound disorder. *Am J Med Genet A,* **164,** 2091-2096.

Peterson, R.L. & Pennington, B.F. (2012) Developmental dyslexia. *Lancet,* **379,** 1997-2007.

Pierron, D., Cortés, N.G., Letellier, T. & Grossman, L.I. (2013) Current relaxation of selection on the human genome: Tolerance of deleterious mutations on olfactory receptors. *Mol Phylogenet Evol,* **66,** 558-564.

Poelmans, G., Buitelaar, J.K., Pauls, D.L. & Franke, B. (2011) A theoretical molecular network for dyslexia: integrating available genetic findings. *Mol Psychiatry*, **16**, 365-382.

Poelmans, G., Engelen, J.J.M., Van Lent-Albrechts, J., Smeets, H.J., Schoenmakers, E., Franke, B., Buitelaar, J.K., Wuisman-Frerker, M., Erens, W., Steyaert, J. & Schrander-Stumpel, C. (2009) Identification of novel dyslexia candidate genes through the analysis of a chromosomal deletion. *Am J Med Genet B Neuropsychiatr Genet*, **150B**, 140-147.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. & Sham, P.C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet,* **81,** 559-575.

Raca, G., Baas, B.S., Kirmani, S., Laffin, J.J., Jackson, C.A., Strand, E.A., Jakielski, K.J. & Shriberg, L.D. (2013) Childhood Apraxia of Speech (CAS) in two patients with 16p11.2 microdeletion syndrome. *Eur J Hum Genet,* **21,** 455-459.

Raskind, W.H., Peter, B., Richards, T.L., Eckert, M.A. & Berninger, V.W. (2013) The genetics of reading disabilities: from phenotypes to candidate genes. *Front Psychol,* **3**.

Roohi, J., Montagna, C., Tegay, D.H., Palmer, L.E., DeVincent, C., Pomeroy, J.C., Christian, S.L., Nowak, N. & Hatchwell, E. (2009) Disruption of contactin 4 in three subjects with autism spectrum disorder. *J Med Genet,* **46,** 176-182.

Rudov, A., Rocchi, M.B.L., Accorsi, A., Spada, G., Procopio, A.D., Olivieri, F., Rippo, M.R. & Albertini, M.C. (2013) Putative miRNAs for the diagnosis of dyslexia, dyspraxia, and specific language impairment. *Epigenetics,* **8,** 1023-1029.

Scerri, T.S., Morris, A.P., Buckingham, L.-L., Newbury, D.F., Miller, L.L., Monaco, A.P., Bishop, D.V.M. & Paracchini, S. (2011) DCDC2, KIAA0319 and CMIP Are Associated with Reading-Related Traits. *Biol Psychiatry,* **70,** 237-245.

Shaywitz, S.E., Shaywitz, B.A., Fletcher, J.M.à. & Escobar, M.D. (1990) Prevalence of reading disability in boys and girls. Results of the Connecticut Longitudinal Study. *JAMA,* **264,** 998–1002.

Simpson, N.H., Ceroni, F., Reader, R.H., Covill, L.E., Knight, J.C., the, S.L.I.C., Hennessy, E.R., Bolton, P.F., Conti-Ramsden, G., O'Hare, A., Baird, G., Fisher, S.E. & Newbury, D.F. (2015) Genome-wide analysis identifies a role for common copy number variants in specific language impairment. Eur J Hum Genet, doi: 10.1038/ejhg.2014.296

Snowling, M., Bishop, D.V. & Stothard, S.E. (2000) Is preschool language impairment a risk factor for dyslexia in adolescence? *J Child Psychol Psychiatry. ,* **41,** 587–600.

Stacklies, W., Redestig, H., Scholz, M., Walther, D. & Selbig, J. (2007) pcaMethods—a bioconductor package providing PCA methods for incomplete data. *Bioinformatics,* **23,** 1164-1167.

Stefansson, H., Meyer-Lindenberg, A., Steinberg, S., Magnusdottir, B., Morgen, K., Arnarsdottir, S., Bjornsdottir, G., Walters, G.B., Jonsdottir, G.A., Doyle, O.M., Tost, H., Grimm, O., Kristjansdottir, S., Snorrason, H., Davidsdottir, S.R., Gudmundsson, L.J., Jonsson, G.F., Stefansdottir, B., Helgadottir, I., Haraldsson, M., Jonsdottir, B., Thygesen, J.H., Schwarz, A.J., Didriksen, M., Stensbol, T.B., Brammer, M., Kapur, S., Halldorsson, J.G., Hreidarsson, S., Saemundsen, E., Sigurdsson, E. & Stefansson, K. (2014) CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature,* **505,** 361-366.

Stefansson, H., Rujescu, D., Cichon, S., Pietilainen, O.P.H., Ingason, A., Steinberg, S., Fossdal, R., Sigurdsson, E., Sigmundsson, T., Buizer-Voskamp, J.E., Hansen, T., Jakobsen, K.D., Muglia, P., Francks, C., Matthews, P.M., Gylfason, A., Halldorsson, B.V., Gudbjartsson, D., Thorgeirsson, T.E., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., Bjornsson, A., Mattiasdottir, S., Blondal, T., Haraldsson, M., Magnusdottir, B.B., Giegling, I., Moller, H.-J., Hartmann, A., Shianna, K.V., Ge, D., Need, A.C., Crombie, C., Fraser, G., Walker, N., Lonnqvist, J., Suvisaari, J., Tuulio-Henriksson, A., Paunio, T., Toulopoulou, T., Bramon, E., Di Forti, M., Murray, R., Ruggeri, M., Vassos, E., Tosato, S., Walshe, M., Li, T., Vasilescu, C., Muhleisen, T.W., Wang, A.G., Ullum, H., Djurovic, S., Melle, I., Olesen, J., Kiemeney, L.A., Franke, B., Sabatti, C., Freimer, N.B., Gulcher, J.R., Thorsteinsdottir, U., Kong, A., Andreassen, O.A., Ophoff, R.A., Georgi, A., Rietschel, M., Werge, T., Petursson, H., Goldstein, D.B., Nothen, M.M., Peltonen, L., Collier, D.A., St Clair, D. & Stefansson, K. (2008) Large recurrent microdeletions associated with schizophrenia. *Nature,* **455,** 232-236.

Sultana, R., Yu, C.-E., Yu, J., Munson, J., Chen, D., Hua, W., Estes, A., Cortes, F., de la Barra, F., Yu, D., Haider, S.T., Trask, B.J., Green, E.D., Raskind, W.H., Disteche, C.M., Wijsman, E., Dawson, G., Storm, D.R., Schellenberg, G.D. & Villacres, E.C. (2002) Identification of a Novel Gene on Chromosome 7q11.2 Interrupted by a Translocation Breakpoint in a Pair of Autistic Twins. *Genomics,* **80,** 129-134.

Taipale, M., Kaminen, N., Nopola-Hemmi, J., Haltia, T., Myllyluoma, B., Lyytinen, H., Muller, K., Kaaranen, M., Lindsberg, P.J., Hannula-Jouppi, K. & Kere, J. (2003) A candidate gene for developmental dyslexia encodes a nuclear tetratricopeptide repeat domain protein dynamically regulated in brain. *Proc Natl Acad Sci USA,* **100,** 11553-11558.

Tammimies, K., Tapia-Paez, I., Ruegg, J., Rosin, G., Kere, J., Gustafsson, J.A. & Nalvarte, I. (2012) The rs3743205 SNP is important for the regulation of the dyslexia candidate gene DYX1C1 by estrogen receptor beta and DNA methylation. *Mol Endocrinol,* **26,** 619-629.

Tarkar, A., Loges, N.T., Slagle, C.E., Francis, R., Dougherty, G.W., Tamayo, J.V., Shook, B., Cantino, M., Schwartz, D., Jahnke, C., Olbrich, H., Werner, C., Raidt, J., Pennekamp, P., Abouhamed, M., Hjeij, R., Kohler, G., Griese, M., Li, Y., Lemke, K., Klena, N., Liu, X., Gabriel, G., Tobita, K., Jaspers, M., Morgan, L.C., Shapiro, A.J., Letteboer, S.J.F., Mans, D.A., Carson, J.L., Leigh, M.W., Wolf, W.E., Chen, S., Lucas, J.S., Onoufriadis, A., Plagnol, V., Schmidts, M., Boldt, K., Uk10K, Roepman, R., Zariwala, M.A., Lo, C.W., Mitchison, H.M., Knowles, M.R., Burdine, R.D., LoTurco, J.J. & Omran, H. (2013) DYX1C1 is required for axonemal dynein assembly and ciliary motility. *Nat Genet,* **45,** 995-1003.

R Core Team (2013) R: A Language and Environment for Statistical Computing. . R Foundation for Statistical Computing, Vienna, Austria.

Tomblin, J.B., Records, N.L., Buckwalter, P., Zhang, X., Smith, E. & O'Brien, M. (1997) Prevalence of specific language impairment in kindergarten children. *J Speech Lang Hear Res,* **40,** 1245-1260.

Turetsky, B.I., Moberg, P.J., Yousem, D.M., Doty, R.L., Arnold, S.E. & Gur, R.E. (2000) Reduced olfactory bulb volume in patients with schizophrenia. *Am J Psychiatry,* **157,** 828-830.

Veerappa, A.M., Saldanha, M., Padakannaya, P. & Ramachandra, N.B. (2013a) Family-based genome-wide copy number scan identifies five new genes of dyslexia involved in dendritic spinal plasticity. *JHum Genet,* **58,** 539-547.

Veerappa, A.M., Saldanha, M., Padakannaya, P. & Ramachandra, N.B. (2013b) Genome-wide copy number scan identifies disruption of PCDH11X in developmental dyslexia. *Am J Med Genet B Neuropsychiatr Genet,* **162B,** 889-897.

Vernes, S.C., Newbury, D.F., Abrahams, B.S., Winchester, L., Nicod, J., Groszer, M., Alarcon, M., Oliver, P.L., Davies, K.E., Geschwind, D.H., Monaco, A.P. & Fisher, S.E. (2008) A functional genetic link between distinct developmental language disorders. *N Engl J Med,* **359,** 2337-2345.

Vernes, S.C., Oliver, P.L., Spiteri, E., Lockstone, H.E., Puliyadi, R., Taylor, J.M., Ho, J., Mombereau, C., Brewer, A., Lowy, E., Nicod, J., Groszer, M., Baban, D., Sahgal, N., Cazier, J.-B., Ragoussis, J., Davies, K.E., Geschwind, D.H. & Fisher, S.E. (2011) Foxp2 Regulates Gene Networks Implicated in Neurite Outgrowth in the Developing Brain. *PLoS Genet,* **7,** e1002145.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F., Hakonarson, H. & Bucan, M. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res,* **17,** 1665-1674.

Wang, K., Zhang, H., Ma, D., Bucan, M., Glessner, J., Abrahams, B., Salyakina, D., Imielinski, M., Bradfield, J., Sleiman, P., Kim, C., Hou, C., Frackelton, E., Chiavacci, R., Takahashi, N., Sakurai, T., Rappaport, E., Lajonchere, C., Munson, J., Estes, A., Korvatska, O., Piven, J., Sonnenblick, L., Alvarez Retuerto, A., Herman, E., Dong, H., Hutman, T., Sigman, M., Ozonoff, S. & Klin, A. (2009) Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature,* **459,** 528 - 533.

Weiss, L., Arking, D., Daly, M. & Chakravarti, A. (2009) A genome-wide linkage and association scan reveals novel loci for autism. *Nature,* **461,** 802 - 808.

Weiss, L.A., Shen, Y., Korn, J.M., Arking, D.E., Miller, D.T., Fossdal, R., Saemundsen, E., Stefansson, H., Ferreira, M.A.R., Green, T., Platt, O.S., Ruderfer, D.M., Walsh, C.A., Altshuler, D., Chakravarti, A., Tanzi, R.E., Stefansson, K., Santangelo, S.L., Gusella, J.F., Sklar, P., Wu, B.-L. & Daly, M.J. (2008)

Association between Microdeletion and Microduplication at 16p11.2 and Autism. *N Engl J Med,* **358,** 667-675.

Whitehouse, A.J., Bishop, D.V., Ang, Q.W., Pennell, C.E. & Fisher, S.E. (2011) CNTNAP2 variants affect early language development in the general population. *Genes Brain Behav,* **10,** 451-456.

Whitehouse, A.J., Mattes, E., Maybery, M.T., Sawyer, M.G., Jacoby, P., Keelan, J.A. & Hickey, M. (2012) Sex-specific associations between umbilical cord blood testosterone levels and language delay in early childhood. *J Child Psychol Psychiatry,* **53,** 726-734.

Willcutt, E.G., Betjemann, R.S., McGrath, L.M., Chhabildas, N.A., Olson, R.K., DeFries, J.C. & Pennington, B.F. (2010) Etiology and neuropsychology of comorbidity between RD and ADHD: the case for multiple-deficit models. *Cortex,* **46,** 1345-1361.

Willcutt, E.G., Pennington, B.F., Olson, R.K., Chhabildas, N. & Hulslander, J. (2005) Neuropsychological analyses of comorbidity between reading disability and attention deficit hyperactivity disorder: in search of the common deficit. *Dev Neuropsychol,* **27,** 35-78.

Willer, C.J., Li, Y. & Abecasis, G.R. (2010) METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics,* **26,** 2190-2191.

Williams, N.M., Franke, B., Mick, E., Anney, R.J., Freitag, C.M., Gill, M., Thapar, A., O'Donovan, M.C., Owen, M.J., Holmans, P., Kent, L., Middleton, F., Zhang-James, Y., Liu, L., Meyer, J., Nguyen, T.T., Romanos, J., Romanos, M., Seitz, C., Renner, T.J., Walitza, S., Warnke, A., Palmason, H., Buitelaar, J., Rommelse, N., Vasquez, A.A., Hawi, Z., Langley, K., Sergeant, J., Steinhausen, H.C., Roeyers, H., Biederman, J., Zaharieva, I., Hakonarson, H., Elia, J., Lionel, A.C., Crosbie, J., Marshall, C.R., Schachar, R., Scherer, S.W., Todorov, A., Smalley, S.L., Loo, S., Nelson, S., Shtir, C., Asherson, P., Reif, A., Lesch, K.P. & Faraone, S.V. (2012) Genome-wide analysis of copy number variants in attention deficit hyperactivity disorder: the role of rare variants and duplications at 15q13.3. *Am J Psychiatry,* **169,** 195-204.

Wiszniewski, W., Hunter, Jill V., Hanchard, Neil A., Willer, Jason R., Shaw, C., Tian, Q., Illner, A., Wang, X., Cheung, Sau W., Patel, A., Campbell, I.M., Gelowani, V., Hixson, P., Ester, Audrey R., Azamian, Mahshid S., Potocki, L., Zapata, G., Hernandez, Patricia P., Ramocki, Melissa B., Santos-Cortez, Regie L., Wang, G., York, Michele K., Justice, M.J., Chu, Zili D., Bader, Patricia I., Omo-Griffith, L., Madduri, Nirupama S., Scharer, G., Crawford, Heather P., Yanatatsaneejit, P., Eifert, A., Kerr, J., Bacino, Carlos A., Franklin, Adiaha I., Goin-Kochel, R.P., Simpson, G., Immken, L., Haque, Muhammad E., Stosic, M., Williams, Misti D., Morgan, Thomas M., Pruthi, S., Omary, R., Boyadjiev, Simeon A., Win, Kay K., Thida, A., Hurles, M., Hibberd, Martin L., Khor, Chiea C., Van Vinh Chau, N., Gallagher, Thomas E., Mutirangura, A., Stankiewicz, P., Beaudet, Arthur L., Maletic-Savatic, M., Rosenfeld, Jill A., Shaffer, Lisa G., Davis, Erica E., Belmont, John W., Dunstan, S., Simmons, C.P., Bonnen, P.E., Leal, Suzanne M., Katsanis, N., Lupski, James R. & Lalani, Seema R. (2013) TM4SF20 Ancestral Deletion and Susceptibility to a Pediatric Disorder of Early Language Delay and Cerebral White Matter Hyperintensities. *Am J Hum Genet*, **93**, 197-210.

Zuko, A., Kleijer, K.T.E., Oguro-Ando, A., Kas, M.J.H., van Daalen, E., van der Zwaag, B. & Burbach, J.P.H. (2013) Contactins in the neurobiology of autism. *Eur J Pharmacol,* **719,** 63-74.

S1: Supplementary Methods

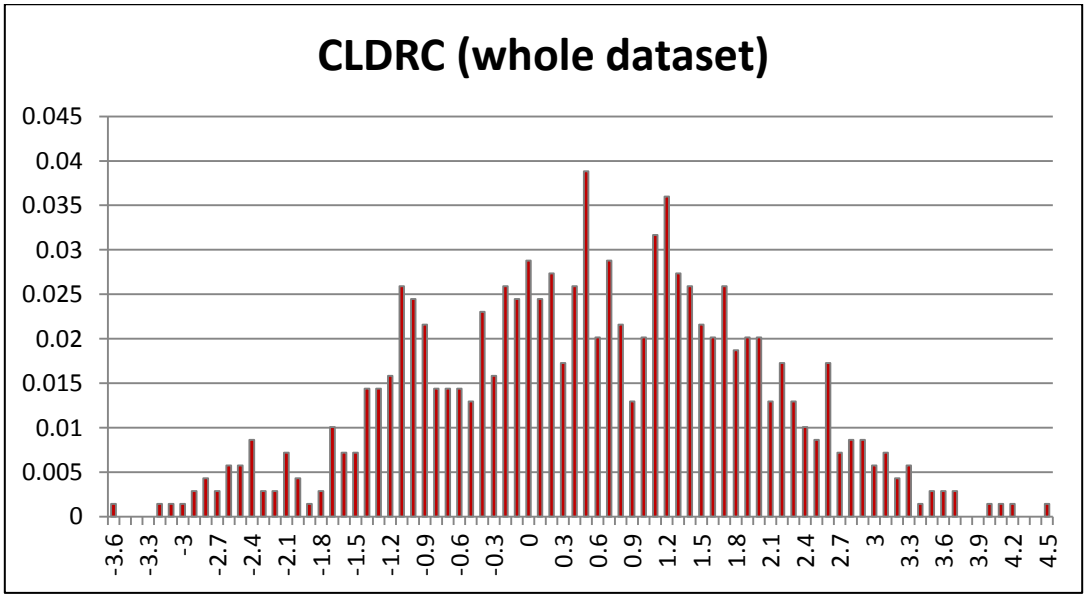
IBG discriminant score

The IBG discriminant score (called *IBGdiscr* hereafter) is a discriminant function empirically developed by John Defries (1985) at the Institute of Behavioral Genetics of University of Colorado at Boulder. This was obtained from the analysis of an independent sample of 140 reading-disabled and 140 control children (DeFries, 1985), to diagnose dyslexia in the context of the CLDRC study. This function is a composite measure of word recognition, spelling and comprehension subtests of the Peabody Individual Achievement Test (Dunn, 1970), as detailed in the formula below:

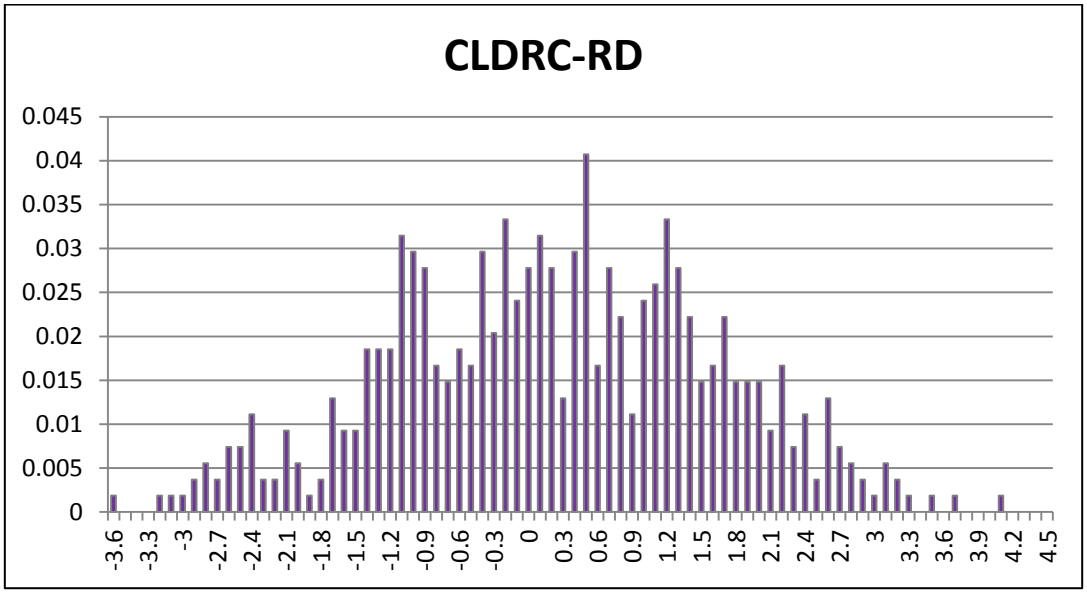
$$IBGdiscr = 1.48121 + 0.078432 * WRead + 0.4810 * WSpell + 0.03453 * WComp,$$

where WRead and WSpell are measures of word reading and spelling (further details in Table S1c of Chapter 2) and WComp is a measure of reading comprehension obtained through a multiple choice test, statistically elaborated in the same way as the other two measures (i.e. adjusted for age and age^2 and standardized against the normative mean of a control population).

a)



b)



c)

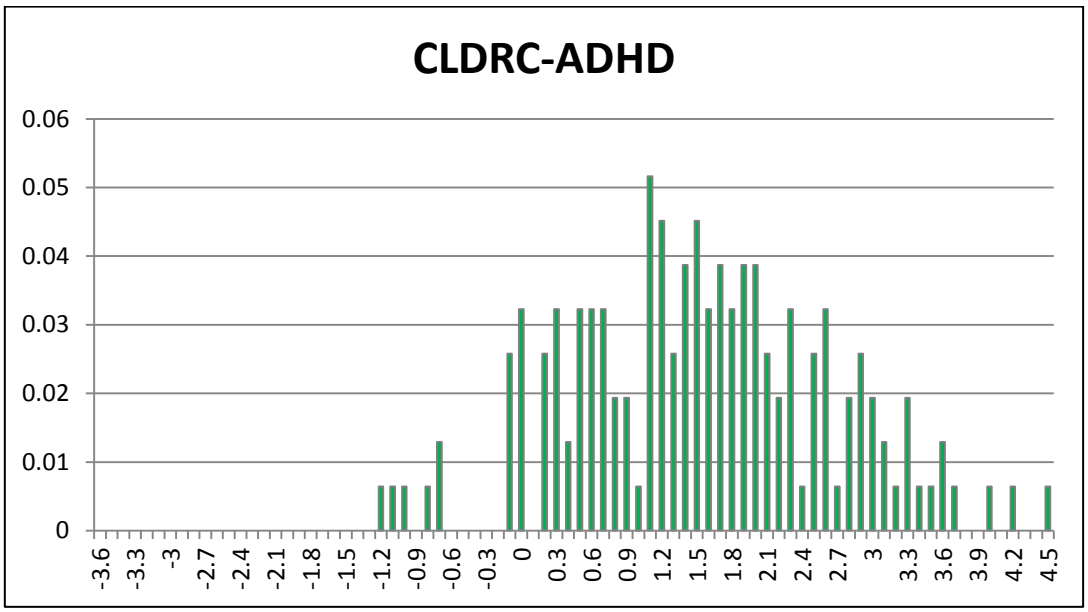


Figure S1. IBG discriminant score distributions in the **a)** CLDRC dataset (N=705 after CNV calling process), **b)** CLDRC-RD (N=546) and **c)** CLDRC-ADHD subset (N=159).

Trait	PC1	IQadjPC1	IBGdiscr
PC1	1	0.94	0.92
IQadjPC1	0.98	1	0.84
IBGdiscr	0.87	0.83	1

Table S1. Pairwise phenotypic correlations of PC1, IQadjPC1 and IBGdiscr in CLDRC-RD (above the diagonal) and CLDRC-ADHD (below the diagonal). These were computed as median Pearson's r coefficients over 100 repeat random samplings of one individual from each unrelated sibship, separately within each subset.

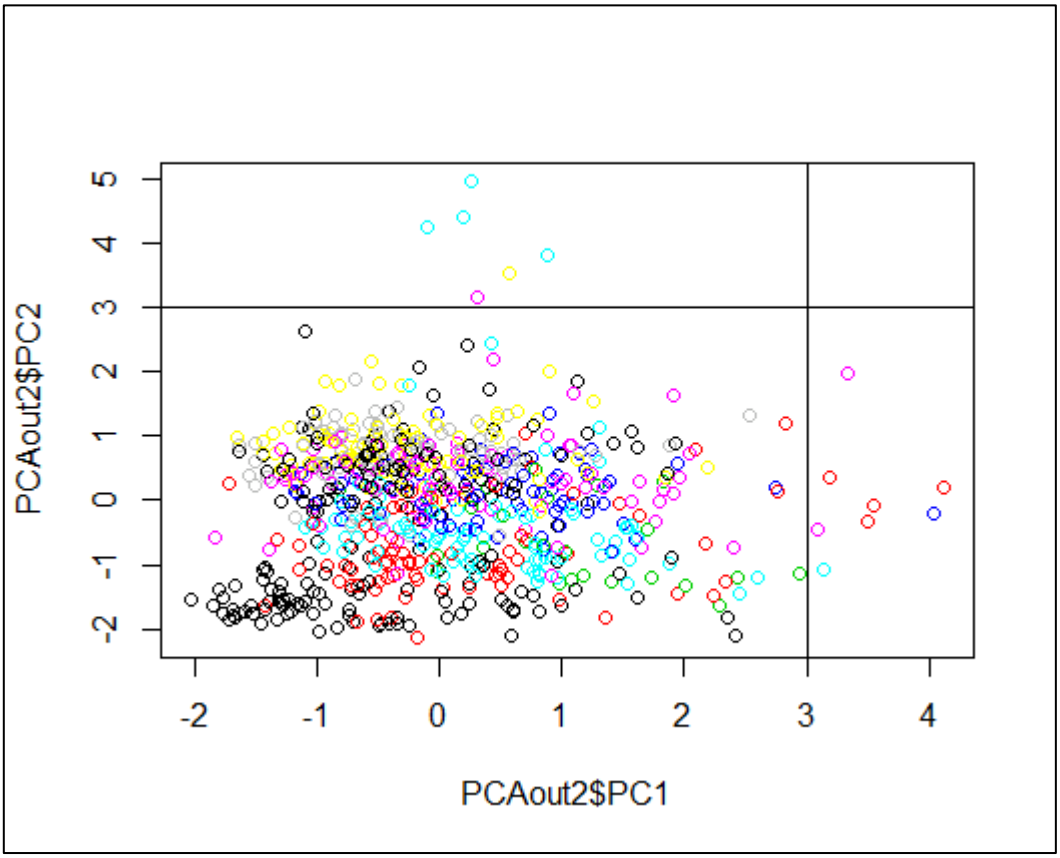


Figure S1d. Scatter plot showing the first two components extracted from the PCA analysis of LRR intensity data of 723,002 SNPs, run on 727 subjects passing genotype and phenotype QC in our previous GWAS meta-analysis (Chapter 3). The first (PC1) and second principal component (PC2) explained 30% and 10% of the total variance in LRR data, while the remaining 98 component scores represented no more than 3% of the total variance each. Samples are colored differently based on the DNA array plate of belonging, in order to detect any potential batch effect among different experiments. Outliers were defined as samples showing scaled PC score >3 for any of the first two principal components, or a PC score >2 for both PC1 and PC2.

S2: CNV calls of interest

Subject	Family	Subset	Chr	Start (kb)	End (kb)	SNPs	Length (kb)	CopyN	Gene	PC1	IQadjPC1	IBGdiscr
IBG145208	4502	ADHD	3	78,923	78,962	5	39	1	ROBO1	-0.48	-0.11	-0.07
IBG143577	4010	RD	15	55,605	55,708	20	103	3	C15orf65,CCPG1,DYX1C1-CCPG1,MIR628,PIGB	-1.73	-1.54	-2.09
IBG143579	4010	RD	15	55,605	55,768	27	163	3	C15orf65,CCPG1,DYX1C1,DYX1C1-CCPG1,MIR628,PIGB	-0.86	-0.59	-1.08
IBG145109	4489	RD	7	146,219	146,389	36	170	1	CNTNAP2	0.28	0.52	0.09
IBG1448951	4442	RD	7	147,117	147,146	5	28	3	CNTNAP2,MIR548I4	-0.97	-1.05	-2.83
IBG144899	4442	RD	7	147,117	147,146	5	28	3	CNTNAP2,MIR548I4	-0.19	-0.11	-0.38
IBG144897	4442	RD	7	147,117	147,146	5	28	3	CNTNAP2,MIR548I4	-0.09	0.24	-0.12

Table S2a. CNVs annotated to candidate susceptibility genes implicated in RD/SLI by previous literature. All the CNVs partially overlapped or encompassed the genes to which they were annotated. All the positions are expressed in hg 19 coordinates.

Subject	Family	Subset	Chr	Start (kb)	End (kb)	SNPs	Length (kb)	CopyN	Gene (distance)[a]	PC1	IQadjPC1	IBGdiscr
IBG144267	4209	RD	22	36,162	36,342	20	181	1	RBFOX2	0.25	0.14	0.69
IBG142597	3002	RD	19	1,377	1,481	34	105	1	APC2,C19orf25,DAZAP1,GAMT,MUM1,NDUFS7,PCSK4,RPS15	-0.37	-0.42	-0.54
IBG113688	5566	RD	19	1,388	1,429	14	41	3	DAZAP1,GAMT,NDUFS7	0.82	1.25	-0.32
IBG112308	3995	RD	19	1,414	1,429	6	16	1	DAZAP1	-0.8	-1.31	-0.15
IBG142178	2526	RD	19	1,425	1,521	39	96	1	ADAMTSL5,APC2,C19orf25,DAZAP1,PCSK4,REEP6,RPS15	0.27	0.11	-0.08
IBG143407	3973	RD	3	21,841	21,852	6	11	1	ZNF385D (48)	-1.26	-1.26	-1.42
IBG142838	3528	ADHD	3	21,841	21,864	8	23	1	ZNF385D (48)	-0.13	0.11	1.45

Table S2b. CNVs annotated to genes in which suggestive associations have been detected in previous GWAS studies of reading and language skills. All the positions are expressed in hg 19 coordinates. [a] When CNVs do not overlap with coding sequences but are located within 50 kb from the 5'- and 3'-UTRs of genes, distance of annotation from these genes is reported in kb in brackets.

S2c)

Subject (sex)[a]	Family	Subset	Chr	Start (kb)	End (kb)	SNPs	Length (kb)	CopyN	Gene (distance)[b]	PC1	IQadjPC1	IBGdiscr
IBG145360	4646	ADHD	7	110,543	110,591	12	48	3	IMMP2L[c]	-0.18	-0.15	0.12
IBG1453652	4646	ADHD	7	110,549	110,591	10	43	3	IMMP2L[c]	0	-0.15	2.17
IBG112228	3960	RD	7	110,769	110,895	27	127	1	IMMP2L[c]	0.94	0.34	1.5
IBG113288	4858	ADHD	7	110,827	110,940	26	112	3	IMMP2L[c]	1.14	1.06	2.56
IBG113289	4858	ADHD	7	110,827	110,940	26	112	3	IMMP2L[c]	1.31	1.42	1.5
IBG113089	4485	ADHD	7	110,840	110,868	8	27	0	IMMP2L[c]	0.14	0.31	1.67
IBG1130861	4485	ADHD	7	110,840	110,931	19	91	1	IMMP2L[c]	-0.71	-0.3	0.62
IBG113089	4485	ADHD	7	110,879	110,933	12	53	1	IMMP2L[c]	0.14	0.31	1.67
IBG112249	3980	RD	7	110,987	111,166	28	179	1	IMMP2L[c]	0.17	0.23	-0.03
IBG145540	5025	RD	7	111,044	111,092	7	48	1	IMMP2L[c]	1.02	1.18	1.31
IBG112259	3984	RD	7	111,054	111,074	4	20	1	IMMP2L[c]	-1.29	-0.96	-1.18
IBG113709	5478	RD	7	111,054	111,200	20	146	1	IMMP2L[c]	-1.5	-1.45	-1.18
IBG112148	3937	RD	7	111,054	111,200	20	146	1	IMMP2L[c]	0.77	0.71	1
IBG112149	3937	RD	7	111,054	111,200	20	146	1	IMMP2L[c]	1.2	1.15	1.83
IBG111229	1631	RD	7	111,108	111,184	13	75	1	IMMP2L[c]	1.32	1.13	1.24
IBG111228	1631	RD	7	111,122	111,184	11	61	1	IMMP2L[c]	-	-	1.96
IBG113569	5413	ADHD	7	111,146	111,235	17	89	1	IMMP2L[c]	-2.06	-1.91	-1.15
IBG113409	5189	RD	7	111,200	111,278	14	79	3	IMMP2L[c]	-0.57	-0.33	-1.39
IBG1128351	4344	ADHD	7	111,235	111,288	12	53	1	IMMP2L (33)[c]	-	-	-
IBG1128352	4344	ADHD	7	111,235	111,288	12	53	1	IMMP2L (33)[c]	-	-	-
IBG1443551	4214	RD	1	71,902	71,970	18	67	3	NEGR1[d]	1.45	1.39	2.87
IBG111949	3523	RD	1	72,455	72,495	13	40	1	NEGR1[d]	1.46	1.11	1.53
IBG145109	4489	RD	7	146,219	146,389	36	170	1	CNTNAP2[d]	0.28	0.52	0.09
IBG1448951	4442	RD	7	147,117	147,146	5	28	3	CNTNAP2,MIR548I4[d]	-0.97	-1.05	-2.83
IBG144899	4442	RD	7	147,117	147,146	5	28	3	CNTNAP2,MIR548I4[d]	-0.19	-0.11	-0.38
IBG144897	4442	RD	7	147,117	147,146	5	28	3	CNTNAP2,MIR548I4[d]	-0.09	0.24	-0.12
IBG113089 (1)	4485	ADHD	X	91,159	91,322	7	163	0	PCDH11X[e,f]	0.14	0.31	1.67
IBG143407 (1)	3973	RD	X	91,175	91,322	5	147	2	PCDH11X[e,f]	-1.26	-1.26	-1.42
IBG113709 (1)	5478	RD	X	91,270	91,335	4	65	2	PCDH11X[e,f]	-1.5	-1.45	-1.18
IBG1137751 (1)	5524	ADHD	X	91,301	91,335	3	34	2	PCDH11X[e,f]	-1.23	-1.26	-0.1
IBG1138751 (1)	5567	RD	X	91,301	91,335	3	34	2	PCDH11X[e,f]	0.51	0.46	0.87
IBG113699 (1)	5468	RD	X	91,301	91,335	3	34	2	PCDH11X[e,f]	0.84	1.03	1.47
IBG1440751 (1)	4177	RD	X	91,301	91,335	3	34	2	PCDH11X[e,f]	0.94	0.86	1.75
IBG112819 (1)	4346	RD	X	91,301	91,335	3	34	2	PCDH11X[e,f]	1.06	1.41	2.24
IBG1444152 (1)	4234	RD	X	91,301	91,335	3	34	2	PCDH11X[e,f]	1.19	1.03	1.59
IBG112309 (1)	3995	RD	X	91,301	91,379	4	78	2	PCDH11X[e,f]	-0.75	-1.27	0.13
IBG1124561 (2)	4153	RD	X	91,752	91,784	4	32	1	PCDH11X[e,f]	0.61	0.53	0.63

Subject (sex)[a]	Family	Subset	Chr	Start (kb)	End (kb)	SNPs	Length (kb)	CopyN	Gene (distance)[b]	PC1	IQadjPC1	IBGdiscr
IBG1452151 (1)	4509	RD	X	91,752	92,357	24	605	2	PCDH11X[e,f]	0.42	0.26	1.27
IBG143778 (2)	4059	RD	X	91,765	91,784	3	18	1	PCDH11X[e,f]	-0.96	-0.85	-0.96
IBG142008 (2)	2426	ADHD	X	91,765	91,784	3	18	1	PCDH11X[e,f]	-0.49	-0.43	0.43
IBG141037 (1)	1472	RD	X	91,765	91,784	3	18	2	PCDH11X[e,f]	0.1	-0.22	1.01
IBG1445751 (1)	4287	RD	X	91,765	91,784	3	18	2	PCDH11X[e,f]	0.37	-0.12	1.33
IBG112258 (2)	3984	RD	X	91,765	91,784	3	18	1	PCDH11X[e,f]	0.84	1.03	1.6
IBG1429261 (2)	3548	RD	X	91,765	91,784	3	18	1	PCDH11X[e,f]	1.27	0.7	2.31
IBG144299 (1)	4211	RD	X	91,784	91,999	7	216	2	PCDH11X[e,f]	1.14	0.63	2.2
IBG1115051 (1)	2142	RD	X	91,784	92,277	18	494	2	PCDH11X[e,f]	1	0.76	1.17
IBG111999 (1)	3541	RD	X	91,870	92,357	18	487	2	PCDH11X[e,f]	0.37	0.37	0.33
IBG112269 (1)	3987	RD	X	91,900	91,999	3	100	2	PCDH11X[e,f]	1.18	0.99	1.35
IBG113369 (1)	5014	RD	X	91,900	92,043	5	144	2	PCDH11X[e,f]	-0.83	-1.24	-1.04
IBG113089 (1)	4485	ADHD	X	91,900	92,043	5	144	2	PCDH11X[e,f]	0.14	0.31	1.67
IBG112539 (1)	4171	RD	X	91,900	92,043	5	144	2	PCDH11X[e,f]	0.36	0.53	0.29
IBG141417 (1)	2153	RD	X	91,900	92,250	12	350	2	PCDH11X[e,f]	-1.84	-2.49	-2.6
IBG141817 (1)	2277	RD	X	91,900	92,250	12	350	2	PCDH11X[e,f]	-1.64	-1.75	-2.17
IBG1122551 (1)	3984	RD	X	91,900	92,277	14	378	2	PCDH11X[e,f]	1.97	1.24	3.45
IBG1448751 (1)	4433	RD	X	91,900	92,357	17	458	2	PCDH11X[e,f]	0.08	0.29	0.78
IBG1129761	4448	ADHD	5	148,869	148,888	9	18	3	CSNK1A1[f]	0.25	0.54	2.13
IBG1434861	3996	ADHD	5	148,869	148,893	13	24	1	CSNK1A1[f]	0.8	0.7	1.96
IBG112748	4324	RD	5	148,874	148,903	11	29	3	CSNK1A1[f]	-	-	-0.45
IBG145349	4644	RD	5	148,874	148,914	12	40	3	CSNK1A1[f]	-0.76	-0.79	-0.95
IBG142519	2877	RD	11	122,456	122,675	103	220	3	UBASH3B[f]	0.76	0.82	1.86
IBG142517	2877	RD	11	122,456	122,675	104	220	3	UBASH3B[f]	-0.47	0.07	-1.24
IBG1425151	2877	RD	11	122,463	122,670	97	208	3	UBASH3B[f]	0	0.05	-0.26
IBG113349	4965	RD	8	10,025	10,121	63	96	1	MSRA[f]	0.07	0.37	0.46
IBG113668	5475	ADHD	7	81,960	82,454	157	493	3	CACNA2D1,PCLO[f]	0.96	1.06	0.76
IBG1124561	4153	RD	13	93,186	93,219	9	33	1	GPC5[f]	0.61	0.53	0.63
IBG143358	3962	RD	2	98,653	98,755	21	102	3	VWA3B[f]	-0.69	-0.37	-0.07
IBG143350	3962	RD	2	98,653	98,759	22	105	3	VWA3B[f]	0.28	0.28	1.35
IBG1122361(2)	3982	RD	X	35,946	36,452	49	505	3	CHDC2,CXorf22,CXorf30,LOC101928564[f]	0.61	0.45	0.99
IBG142188	2691	RD	22	18,887	21,464	751	2,577	3	69 genes[f,h]	0.16	0.22	-0.1
IBG112399	4050	RD	2	228,241	228,258	9	18	3	TM4SF20[g]	-0.71	-1.11	-1.18
IBG112599	4197	RD	2	228,243	228,258	8	15	3	TM4SF20[g]	-0.32	0	-1.08
IBG1131261	4503	RD	2	228,243	228,258	8	15	3	TM4SF20[g]	-0.07	-0.31	-0.68
IBG112828	4336	ADHD	2	228,243	228,258	8	15	3	TM4SF20[g]	-0.05	0.06	1.71
IBG112548	4170	RD	2	228,243	228,258	8	15	3	TM4SF20[g]	0.31	0.23	0.5
IBG141977	2379	RD	2	228,243	228,258	8	15	3	TM4SF20[g]	0.6	0.49	1.1

Subject (sex)[a]	Family	Subset	Chr	Start (kb)	End (kb)	SNPs	Length (kb)	CopyN	Gene (distance)[b]	PC1	IQadjPC1	IBGdiscr
IBG1133351	4954	RD	2	228,243	228,258	8	15	3	TM4SF20[g]	0.66	0.52	1.21
IBG143059	3588	ADHD	2	228,243	228,258	8	15	3	TM4SF20[g]	0.73	0.63	2.5
IBG145150	4495	RD	2	228,243	228,258	8	15	3	TM4SF20[g]	1.38	1.46	2.03
IBG1448951	4442	RD	2	228,244	228,258	7	14	3	TM4SF20[g]	-0.97	-1.05	-2.83
IBG112319	4000	RD	2	228,244	228,258	7	14	3	TM4SF20[g]	0.27	0.65	0.79
IBG1438151	4063	RD	2	228,244	228,258	7	14	3	TM4SF20[g]	0.77	0.53	1.7
IBG143819	4063	RD	2	228,244	228,258	7	14	3	TM4SF20[g]	0.82	1.01	1.21
IBG1415951	2352	RD	2	228,244	228,258	7	14	3	TM4SF20[g]	1.24	0.34	2.12
IBG141590	2352	RD	2	228,245	228,258	5	13	3	TM4SF20 (1)[g]	1.55	0.47	0.89

Table S2c. CNVs annotated to genes previously identified in CNV studies of RD and/or SLI. When a CNV is annotated to more than ten RefSeq genes, these are reported in a footnote (see below). All the positions are expressed in hg 19 coordinates.

[a] For CNVs called in chromosome X, sex information on the subjects is reported in brackets ("1" = male; "2" = female). [b] When CNVs do not overlap with coding sequences but are located within 50 kb from the 5'- and 3'-UTRs of genes, distance of annotation from these genes is reported in kb in brackets. [c] Pagnamenta et al. (2010), Girirajan et al. (2011). [d] Veerappa et al. (2013a). [e] Veerappa et al. (2013b). [f] Simpson et al. (2015). [g] Wisznieski et al. (2013). [h] Genes encompassed: AIFM3, ARVCF, BCRP2, C22orf29, C22orf39, CDC45, CLDN5, CLTCL1, COMT, CRKL, DGCR10, DGCR11, DGCR14, DGCR2, DGCR5, DGCR6, DGCR6L, DGCR8, DGCR9, GNB1L, GP1BB, GSC2, HIRA, KLHL22, LINC00895, LINC00896, LOC100652736, LOC284865, LOC388849, LOC400891, LOC729444, LZTR1, MED15, MIR1286, MIR1306, MIR185, MIR3618, MIR4761, MIR6816, MRPL40, P2RX6, P2RX6P, PI4KA, PI4KAP1, POM121L4P, PRODH, RANBP1, RIMBP3, RTN4R, SCARF2, SEPT5, SEPT5-GP1BB, SERPIND1, SLC25A1, SLC7A4, SNAP29, TANGO2, TBX1, THAP7, THAP7-AS1, TMEM191A, TMEM191B, TRMT2A, TSSK2, TUBA3FP, TXNRD2, UFD1L, ZDHHC8, ZNF74.

Subject	Family	Subset	Chr	Start (kb)	End (kb)	SNPs	Length (kb)	CopyN	Gene	PC1	IQadjPC1	IBGdiscr
IBG1128351	4344	ADHD	15	22,750	23,273	113	522	3	CYFIP1,GOLGA8I,LOC283683,NIPA1, NIPA2,TUBGCP5,WHAMMP3[a]	-	-	-
IBG112838	4344	ADHD	15	22,750	23,273	113	522	3	CYFIP1,GOLGA8I,LOC283683, NIPA1,NIPA2,TUBGCP5,WHAMMP3[a]	-	-	-
IBG1128352	4344	ADHD	15	22,750	23,273	113	522	3	CYFIP1,GOLGA8I,LOC283683, NIPA1,NIPA2,TUBGCP5,WHAMMP3[a]	-	-	-
IBG112839	4344	ADHD	15	22,750	23,273	113	522	3	CYFIP1,GOLGA8I,LOC283683, NIPA1,NIPA2,TUBGCP5,WHAMMP3[a]	-	-	-
IBG113439	5247	ADHD	16	29,595	30,198	116	603	3	29 genes[b]	0.55	0.97	1.76
IBG143568	4013	RD	17	14,101	15,345	617	1,244	3	11 genes[c]	0.24	0.73	-0.18
IBG145109	4489	RD	16	14,930	16,303	420	1,374	3	27 genes[d]	0.28	0.52	0.09
IBG111948	3523	RD	16	14,975	16,303	419	1,328	3	27 genes[d]	-1.55	-1.53	-1.61
IBG142188	2691	RD	22	18,887	21,464	751	2,577	3	69 genes[e]	0.16	0.22	-0.1

Table S2d. Annotated CNVs which showed large/complete overlap with canonical neuropsychiatric CNVs assessed in Stefansson et al. (2014). All the CNVs partially overlapped or encompassed the genes to which they were annotated. When a CNV is annotated to more than ten RefSeq genes, these are reported in a footnote (see below). All the positions are expressed in hg 19 coordinates.

[a] 15q11.2 is a susceptibility region for several neurological dysfunctions, including language delays. CNVs in this region are also associated with schizophrenia (SCZ).

[b] Duplications/deletions in 16p11.2 (29.5-30.2 Mb) are involved in Developmental Delay (DD), Autism Spectrum Disorders (ASD), Intellectual Disability (ID) and SCZ. Genes encompassed: ALDOA, ASPHD1, C16orf54, C16orf92, CDIPT, CDIPT-AS1, CORO1A, DOC2A, FAM57B, GDPD3, HIRIP3, INO80E, KCTD13, KIF22, MAPK3, MAZ, MVP, PAGR1, PPP4C, PRRT2, QPRT, SEZ6L2, SLC7A5P1, SPN, TAOK2, TBX6, TMEM219, YPEL3, ZG16.

[c] Deletions in 17p12 are involved in ASD and SCZ. Genes encompassed: DCDRT15, CDRT4, CDRT7, CDRT8, COX10, HS3ST3B1, MGC12916, MIR4731, PMP22, TEKT3, TVP23C-CDRT4.

[d] Duplications/deletions in 16p13.11 are involved in DD, ASD and SCZ. Genes encompassed: ABCC1, ABCC6, C16orf45, FOPNL, KIAA0430, LOC100288162, MIR3179-1, MIR3179-2, MIR3179-3, MIR3180-1, MIR3180-2, MIR3180-3, MIR3180-4, MIR484, MIR6506, MIR6511A-2, MIR6511B-1, MIR6770-2, MPV17L, MYH11, NDE1, NOMO1, NPIPA1, NPIPA5, NTAN1, PDXDC1, RRN3.

[e] 22q11.21 is the critical region of Digeorge Syndrome. Deletions in this region are also implicated in SCZ. Genes encompassed: AIFM3, ARVCF, BCRP2, C22orf29, C22orf39, CDC45, CLDN5, CLTCL1, COMT, CRKL, DGCR10, DGCR11, DGCR14, DGCR2, DGCR5, DGCR6, DGCR6L, DGCR8, DGCR9, GNB1L, GP1BB, GSC2, HIRA, KLHL22, LINC00895, LINC00896, LOC100652736, LOC284865, LOC388849, LOC400891, LOC729444, LZTR1, MED15, MIR1286, MIR1306, MIR185, MIR3618, MIR4761, MIR6816, MRPL40, P2RX6, P2RX6P, PI4KA, PI4KAP1, POM121L4P, PRODH, RANBP1, RIMBP3, RTN4R, SCARF2, SEPT5, SEPT5-GP1BB, SERPIND1, SLC25A1, SLC7A4, SNAP29, TANGO2, TBX1, THAP7, THAP7-AS1, TMEM191A, TMEM191B, TRMT2A, TSSK2, TUBA3FP, TXNRD2, UFD1L, ZDHHC8, ZNF74.

Subject	Family	Subset	Chr	Start (kb)	End (kb)	SNPs	Length (kb)	CopyN	Gene	PC1	IQadjPC1	IBGdiscr
IBG145848	5474	RD	15	31,964	32,514	109	551	1	CHRNA7	-0.01	0.64	-0.36
IBG142928	3548	RD	15	32,019	32,514	108	496	3	CHRNA7	-0.13	-0.37	-0.03
IBG143639	4032	RD	15	32,019	32,514	108	496	3	CHRNA7	0.87	0.44	1.8
IBG1429261	3548	RD	15	32,019	32,514	108	496	3	CHRNA7	1.27	0.7	2.31
IBG142920	3548	RD	15	32,020	32,514	107	494	3	CHRNA7	1.47	1.77	1.12
IBG112739	4309	RD	15	32,020	32,514	107	494	3	CHRNA7	-	-	2.69
IBG1443551	4214	RD	15	32,049	32,514	103	465	3	CHRNA7	1.45	1.39	2.87
IBG142478	2862	RD	15	32,061	32,514	100	453	3	CHRNA7	0.34	0.07	1.23
IBG113029	4478	RD	15	32,380	32,514	25	134	3	CHRNA7	0.24	0.24	0.75
IBG113028	4478	RD	15	32,380	32,514	25	134	3	CHRNA7	1.21	0.81	2.17

Table S2e. CNVs annotated to *CHRNA7* (15q13.3). Deletions in this region have been found in ASD, SCZ and DD. All the CNVs partially overlapped or encompassed *CHRNA7*. All the positions are expressed in hg 19 coordinates.

S2f)

Subject	Family	Subset	Chr	Start (kb)	End (kb)	SNPs	Length (kb)	CopyN	Gene (distance)[a]	PC1	IQadjPC1	IBGdiscr
IBG143618	4021	RD	19	20,626	20,708	12	81	1	ZNF737 (13)	-0.84	-0.54	-1.48
IBG113528	5374	ADHD	19	20,626	20,708	12	81	1	ZNF737 (13)	-0.48	-0.38	1.67
IBG112919	4418	RD	19	20,626	20,708	12	81	1	ZNF737 (13)	-0.37	0.11	-0.2
IBG112599	4197	RD	19	20,626	20,708	12	81	1	ZNF737 (13)	-0.32	0	-1.08
IBG144330	4216	RD	19	20,626	20,708	12	81	1	ZNF737 (13)	-0.27	-0.24	0.11
IBG143388	3972	RD	19	20,626	20,708	12	81	1	ZNF737 (13)	-0.04	-0.09	-0.37
IBG144319	4213	RD	19	20,626	20,708	12	81	1	ZNF737 (13)	0.5	0.69	1.03
IBG1439261	4147	RD	19	20,626	20,708	12	81	1	ZNF737 (13)	0.79	0.22	1.99
IBG112088	3904	RD	19	20,626	20,708	12	81	1	ZNF737 (13)	1.8	1.09	1.76
IBG140227	1200	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	-1.78	-1.45	-3
IBG112528	4167	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	-1.4	-1.48	-1.38
IBG142457	2867	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	-1.19	-0.86	-1.28
IBG1449162	4445	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	-1.11	-1.14	-1.63
IBG142789	3518	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	-1.09	-1.25	-0.85
IBG1448951	4442	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	-0.97	-1.05	-2.83
IBG141237	1685	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	-0.95	-1.06	-1.53
IBG144917	4445	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	-0.91	-0.59	-1.1
IBG1453351	4657	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	-0.89	-1.14	-1.28
IBG142290	2690	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	-0.81	-0.7	-1.19
IBG144478	4255	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	-0.74	-0.93	-0.42
IBG145717	5275	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	-0.73	-0.73	-1.06
IBG1117652	2799	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	-0.7	-0.97	-0.42
IBG1125761	4183	ADHD	19	20,626	20,715	13	89	1	ZNF737 (6)	-0.69	-0.52	0.61
IBG1448861	4436	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	-0.69	-0.45	-0.8
IBG144447	4239	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	-0.58	-0.28	-1.21
IBG1453651	4646	ADHD	19	20,626	20,715	13	89	1	ZNF737 (6)	-0.4	-0.62	1.37
IBG112498	4164	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	-0.38	-0.7	0.12
IBG113088	4485	ADHD	19	20,626	20,715	13	89	1	ZNF737 (6)	-0.31	-0.36	0.33

Subject	Family	Subset	Chr	Start (kb)	End (kb)	SNPs	Length (kb)	CopyN	Gene (distance)[a]	PC1	IQadjPC1	IBGdiscr
IBG113389	5040	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	-0.28	-0.12	-1.13
IBG143457	3999	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	-0.2	-0.2	-0.3
IBG144899	4442	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	-0.19	-0.11	-0.38
IBG145360	4646	ADHD	19	20,626	20,715	13	89	1	ZNF737 (6)	-0.18	-0.15	0.12
IBG145368	4646	ADHD	19	20,626	20,715	13	89	1	ZNF737 (6)	-0.16	-0.23	1.27
IBG144308	4210	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	-0.13	0.11	-0.55
IBG144897	4442	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	-0.09	0.24	-0.12
IBG1126361	4218	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	0	-0.17	-0.13
IBG1453652	4646	ADHD	19	20,626	20,715	13	89	1	ZNF737 (6)	0	-0.15	2.17
IBG145188	4498	ADHD	19	20,626	20,715	13	89	1	ZNF737 (6)	0.06	0.42	1.17
IBG113089	4485	ADHD	19	20,626	20,715	13	89	1	ZNF737 (6)	0.14	0.31	1.67
IBG113118	4579	ADHD	19	20,626	20,715	13	89	1	ZNF737 (6)	0.16	-0.09	2.96
IBG113388	5040	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	0.21	-0.28	0.85
IBG145109	4489	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	0.28	0.52	0.09
IBG140229	1200	RD	19	20,626	20,715	13	89	0	ZNF737 (6)	0.28	0.34	0.06
IBG1427861	3518	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	0.29	-0.2	0.9
IBG113728	5507	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	0.3	0.65	0.74
IBG113899	5569	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	0.35	0.32	0.61
IBG112459	4153	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	0.39	0.04	0.92
IBG145107	4489	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	0.43	1.02	-0.19
IBG112799	4332	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	0.47	0.88	0.79
IBG143077	3598	ADHD	19	20,626	20,715	13	89	0	ZNF737 (6)	0.54	0.59	1.74
IBG1457151	5275	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	0.58	0.2	1.47
IBG1124561	4153	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	0.61	0.53	0.63
IBG1443361	4216	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	0.61	0.53	1.25
IBG112089	3904	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	0.74	0.58	1.32
IBG1434561	3999	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	0.8	0.53	1.09
IBG112258	3984	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	0.84	1.03	1.6

Subject	Family	Subset	Chr	Start (kb)	End (kb)	SNPs	Length (kb)	CopyN	Gene (distance)[a]	PC1	IQadjPC1	IBGdiscr
IBG143380	3972	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	0.86	1.24	1.99
IBG144840	4430	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	0.9	1.09	1.25
IBG1444751	4255	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	0.99	0.07	2.37
IBG1444451	4239	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	1.09	1.14	1.92
IBG1125951	4197	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	1.11	1.21	2.28
IBG113288	4858	ADHD	19	20,626	20,715	13	89	1	ZNF737 (6)	1.14	1.06	2.56
IBG143610	4021	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	1.21	0.24	2.58
IBG145320	4589	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	1.21	1.7	0.25
IBG143858	4089	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	1.31	1.41	2.61
IBG142459	2867	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	1.52	1.44	2.11
IBG1120851	3904	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	1.56	1.08	2.7
IBG142567	2963	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	1.61	1.93	1.71
IBG112818	4346	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	1.74	1.64	2.04
IBG1443051	4210	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	1.77	1.18	1.91
IBG143850	4089	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	1.79	1.73	2.45
IBG1122551	3984	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	1.97	1.24	3.45
IBG1455761	5041	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	1.99	1.89	2.12
IBG144957	4453	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	2.09	2.14	2.58
IBG112478	4154	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	2.13	2.4	2.05
IBG112959	4437	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	2.17	2.15	2.54
IBG111768	2799	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	-	-	0.98
IBG111928	3519	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	-	-	-0.11
IBG111769	2799	RD	19	20,626	20,715	13	89	1	ZNF737 (6)	-	-	-0.9
IBG111828	2856	RD	19	20,630	20,715	12	85	1	ZNF737 (6)	-0.43	-0.13	-0.52

Table S2f. CNV calls annotated to *ZNF737* (19p12). All the positions are expressed in hg 19 coordinates. [a] When CNVs do not overlap with the coding sequence but are located within 50 kb from the 5'- and 3'-UTRs of *ZNF737*, distance of annotation from the gene is reported in kb in brackets.

S3: Supplementary Results

Chr	SNP	Position (hg19)	Zscore	P-value	Direction[a]	HetPVal[b]	Gene (distance)[c]
5	rs283107	32101400	2.89	0.004	++	0.6	PDZD2(0)|GOLPH3(+23.42)
5	rs2468506	36449552	2.99	0.003	++	0.79	
5	rs10061999	36450102	2.94	0.003	++	0.84	
5	rs7709504	36450612	2.98	0.003	++	0.82	
5	rs2468509	36450876	2.99	0.003	++	0.76	
5	rs17286376	36453541	2.86	0.004	++	0.82	
5	rs2455274	36456056	2.9	0.004	++	0.83	
5	rs2455275	36456308	2.94	0.003	++	0.84	
5	rs7730299	36460107	3.03	0.002	++	0.76	
5	rs2455280	36460425	2.95	0.003	++	0.85	
5	rs17358533	36460462	3.0	0.003	++	0.77	
5	rs2468519	36461331	2.99	0.003	++	0.81	

Table S3a. Top associated probes ($p < 0.005$) in the GWAS meta-analysis of PC1 with CNV state (implemented in PLINK QFAM). Genome-wide significance threshold: $\alpha = 7.6 \times 10^{-6}$ (corrected for multiple testing of ~6,586 SNPs encompassed by at least one putative CNV event in both CLDRC-RD and CLDRC-ADHD). Probes are ordered by chromosome and position to facilitate the interpretation of results in terms of consecutive probes associated with PC1.
[a] The direction of effect refers to the "CNV+" state (i.e. copy number other than 2) and is reported for subsets in the following order: CLDRC-RD, CLDRC-ADHD. [b] Test for the homogeneity of effect sizes across the different subsets ($p \geq 0.05$ indicates homogeneous effects). [c] Physical distance (kb) from close genes (in a ±50kb range from each marker) is indicated, along with orientation based on the direction of transcription ("-" = upstream of 5'-UTR, "+" = downstream of 3'-UTR).

S3b)

Chr	SNP	Position (hg19)	Zscore	P-value	Direction[a]	HetPVal[b]	Gene (distance)[c]
3	rs1479546	2667189	2.94	0.003	++	0.31	CNTN4(0)
3	rs6803088	2668594	2.92	0.003	++	0.27	CNTN4(0)
3	rs2600318	2670624	2.96	0.003	++	0.34	CNTN4(0)
5	rs283107	32101400	3.07	0.002	++	0.74	PDZD2(0)|GOLPH3(+23.42)
5	rs2468506	36449552	3.03	0.002	++	0.76	
5	rs10061999	36450102	2.90	0.004	++	0.74	
5	rs2468509	36450876	2.85	0.004	++	0.81	
5	rs17286376	36453541	2.89	0.004	++	0.78	
5	rs2455274	36456056	2.86	0.004	++	0.80	
5	rs2455275	36456308	2.85	0.004	++	0.77	
5	rs7730299	36460107	2.90	0.004	++	0.80	
5	rs2455280	36460425	2.90	0.004	++	0.76	
5	rs17358533	36460462	2.93	0.003	++	0.81	
5	rs2468519	36461331	2.82	0.005[d]	++	0.82	
6	rs7751205	168579302	-3.30	9.7×10^{-4}	--	0.75	
6	rs4708445	168580694	-3.21	0.001	--	0.79	
6	rs11960954	168580741	-3.33	8.8×10^{-4}	--	0.75	
6	rs9455968	168581362	-3.39	7.1×10^{-4}	--	0.77	
6	rs9455971	168582182	-3.26	0.001	--	0.75	
6	rs9455973	168583006	-3.27	0.001	--	0.81	
6	rs2880102	168583032	-3.22	0.001	--	0.73	
6	rs9355178	168589242	-3.16	0.002	--	0.71	
6	rs9283861	168592134	-3.28	0.001	--	0.78	
6	rs12198918	168593739	-3.09	0.002	--	0.74	
6	rs9346533	168593956	-3.07	0.002	--	0.72	
6	rs12213783	168595832	-3.18	0.001	--	0.79	
10	rs7095004	68223696	-2.88	0.004	--	0.86	CTNNA3(0)
10	rs10822834	68224205	-2.84	0.005[d]	--	0.88	CTNNA3(0)
10	rs4745900	68224593	-2.88	0.004	--	0.84	CTNNA3(0)
10	rs2441727	68224886	-2.85	0.004	--	0.89	CTNNA3(0)
10	rs12220315	68225548	-2.92	0.004	--	0.90	CTNNA3(0)
10	rs4587626	68230347	-2.91	0.004	--	0.88	CTNNA3(0)
10	rs12249344	68231810	-2.92	0.004	--	0.91	CTNNA3(0)
10	rs11817581	68237143	-2.83	0.005	--	0.91	CTNNA3(0)
10	rs10822837	68242672	-2.86	0.004	--	0.86	CTNNA3(0)
11	rs4537777	55241556	-2.82	0.005[d]	--	0.81	
11	rs534345	55256498	-2.92	0.003	--	0.75	
11	rs17158615	55258370	-3.02	0.003	--	0.75	

Chr	SNP	Position (hg19)	Zscore	P-value	Direction[a]	HetPVal[b]	Gene (distance)[c]
11	rs10896971	55264310	-2.93	0.003	--	0.70	
11	rs12272148	55272791	-2.90	0.004	--	0.80	OR4C15(-48.99)
11	rs559362	55275456	-2.98	0.003	--	0.77	OR4C15(-46.33)
11	rs12417844	55282064	-3.00	0.003	--	0.76	OR4C15(-39.72)
11	rs17159005	55303865	-3.02	0.003	--	0.75	OR4C16(-35.74)|OR4C15(-17.92)
11	rs526821	55306151	-2.91	0.004	--	0.82	OR4C16(-33.45)|OR4C15(-15.63)
11	rs17580938	55311980	-2.93	0.003	--	0.81	OR4C16(-27.62)|OR4C15(-9.802)
11	rs504661	55312683	-3.02	0.003	--	0.77	OR4C16(-26.92)|OR4C15(-9.099)
11	rs17581191	55316023	-2.87	0.004	--	0.70	OR4C16(-23.58)|OR4C15(-5.759)
11	rs509882	55321055	-3.07	0.002	--	0.74	OR4C16(-18.55)|OR4C15(-0.727)|OR4C11(+49.86)
11	rs17496724	55322099	-2.97	0.003	--	0.76	OR4C16(-17.5)|OR4C15(0)|OR4C11(+48.82)
11	rs12790125	55322539	-2.88	0.004	--	0.76	OR4C16(-17.06)|OR4C15(0)|OR4C11+48.38)
11	rs17581700	55322606	-3.03	0.002	--	0.72	OR4C16(-17)|OR4C15(0)|OR4C11(+48.31)
11	rs12225462	55322638	-3.05	0.002	--	0.73	OR4C16(-16.96)|OR4C15(0)|OR4C11(+48.28)
11	rs506988	55325928	-2.98	0.003	--	0.78	OR4C16(-13.68)|OR4C15(+3.033)|OR4C11(+44.99)
11	rs1394428	55335878	-3.09	0.002	--	0.73	OR4C16(-3.725)|OR4C15(+12.98)|OR4C11(+35.04)
11	rs1459101	55339652	-2.89	0.004	--	0.73	OR4C16(0)|OR4C15(+16.76)|OR4C11(+31.26)
11	rs558465	55339748	-2.97	0.003	--	0.74	OR4C16(0)|OR4C15(+16.85)|OR4C11(+31.17)
11	rs557590	55339829	-2.93	0.003	--	0.78	OR4C16(0)|OR4C15(+16.93)|OR4C11(+31.09)
11	rs559449	55340379	-2.90	0.004	--	0.84	OR4C16(0)|OR4C15(+17.48)|OR4C11(+30.54)
11	rs35992551	55340631	-3.00	0.003	--	0.76	OR4C16(+0.095)|OR4C15(+17.74)|OR4C11(+30.29)
11	rs12421826	55343036	-3.01	0.003	--	0.77	OR4C16(+2.5)|OR4C15(+20.14)|OR4C11(+27.88)
11	rs2903854	55360213	-2.94	0.003	--	0.77	OR4P4(-45.62)|OR4C16(+19.68)|OR4C15(+37.32)|OR4C11(+10.7)
11	rs546140	55361808	-2.94	0.003	--	0.79	OR4P4(-44.02)|OR4C16(+21.27)|OR4C15(+38.91)|OR4C11(+9.108)
11	rs578686	55362955	-2.93	0.003	--	0.74	OR4P4(-42.88)|OR4C16(+22.42)|OR4C15(+40.06)|OR4C11(+7.961)

Table S3b. Top associated probes ($p < 0.005$) in the GWAS meta-analysis of IQ-adjusted PC1 with CNV state (implemented in PLINK QFAM). Genome-wide significance threshold: $\alpha = 7.6 \times 10^{-6}$ (corrected for multiple testing of ~6,586 SNPs encompassed by at least one putative CNV event in both CLDRC-RD and CLDRC-ADHD). Probes are ordered by chromosome and position to facilitate the interpretation of results in terms of consecutive probes associated with IQ-adjusted PC1.
[a] The direction of effect refers to the "CNV+" state (i.e. copy number other than 2) and is reported for subsets in the following order: CLDRC-RD, CLDRC-ADHD. [b] Test for the homogeneity of effect sizes across the different subsets ($p \geq 0.05$ indicates homogeneous effects). [c] Physical distance (kb) from close genes (in a ±50kb range from each marker) is indicated, along with orientation based on the direction of transcription ("-" = upstream of 5'-UTR, "+" = downstream of 3'-UTR). [d] Actual p-value < 0.005 (rounded to the third decimal place).

Chr	SNP	Position (hg19)	Zscore	P-value	Direction[a]	HetPVal[b]	Gene (distance)[c]
7	rs7779972	138746752	4.61	4×10^{-6}	++	0.92	ZC3HAV1(0)|ZC3HAV1L(-25.98)
17	rs6502435	15072464	4.52	6.1×10^{-6}	++	0.85	
18	rs11876036	66747568	-4.24	2.3×10^{-5}	--	0.69	CCDC102B(+25.14)
2	rs6761959	78714978	-4.14	3.5×10^{-5}	--	0.02	
10	rs7916256	55486017	4.13	3.7×10^{-5}	++	0.97	
8	rs17634977	17780036	4.11	4×10^{-5}	++	1	PCM1(-0.329)|FGL1(-26.99)
8	rs7844572	140349351	4.07	4.8×10^{-5}	++	0.76	
13	rs9578596	24002382	-4.04	5.3×10^{-5}	--	0.26	SACS(0)
10	rs10904254	4480571	-4.04	5.3×10^{-5}	--	0.15	
6	rs12528232	44982593	-3.92	8.7×10^{-5}	--	0.58	SUPT3H(0)
6	rs9467759	26464472	3.91	9.4×10^{-5}	++	0.86	BTN3A3(+10.83)|BTN3A1(+49.03)|BTN2A1(0)|BTN1A1(-37.02)
16	rs9930322	17417219	3.9	9.7×10^{-5}	++	0.21	XYLT1(0)

Table S3c. Top associated probes ($p < 1x10^{-4}$) in the GWAS meta-analysis of PC1 with probe intensity data (implemented in FamCNV). Genome-wide significance threshold: $\alpha = 7.1x10^{-8}$, corrected for multiple testing of 704,855 autosomal probes.

[a] The direction of effect refers to the rho correlation coefficient between the LRR intensity signal and PC1 and is reported for subsets in the following order: CLDRC-RD, CLDRC-ADHD. [b] Test for the homogeneity of effect sizes across the different subsets ($p \geq 0.05$ indicates homogeneous effects). [c] Physical distance (kb) from close genes (in a ±50kb range from each marker) is indicated, along with orientation based on the direction of transcription ("-" = upstream of 5'-UTR, "+" = downstream of 3'-UTR).

Chr	SNP	Position (hg19)	Zscore	P-value	Direction[a]	HetPVal[b]	Gene (distance)[c]
10	rs7916256	55486017	4.61	4.1×10^{-6}	++	0.86	
17	rs6502435	15072464	4.52	6.3×10^{-6}	++	0.86	
7	rs7779972	138746752	4.24	2.3×10^{-5}	++	0.98	ZC3HAV1(0)|ZC3HAV1L(-25.98)
2	rs6761959	78714978	-4.02	5.9×10^{-5}	--	0.04	
2	rs3115027	133302257	4.01	6×10^{-5}	++	0.91	GPR39(0)
13	rs9578596	24002382	-4.01	6.2×10^{-5}	--	0.29	SACS(0)
15	rs4260008	86016573	3.97	7.3×10^{-5}	++	0.09	AKAP13(0)

Table S3d. Top associated probes ($p < 1x10^{-4}$) in the GWAS meta-analysis of IQ-adjusted PC1 with probe intensity data (implemented in FamCNV). Genome-wide significance threshold: $\alpha = 7.1x10^{-8}$, corrected for multiple testing of 704,855 autosomal probes. [a] The direction of effect refers to the rho correlation coefficient between the LRR intensity signal and IQ-adjusted PC1 and is reported for subsets in the following order: CLDRC-RD, CLDRC-ADHD. [b] Test for the homogeneity of effect sizes across the different subsets ($p \geq 0.05$ indicates homogeneous effects). [c] Physical distance (kb) from close genes (in a ±50kb range from each marker) is indicated, along with orientation based on the direction of transcription ("-" = upstream of 5'-UTR, "+" = downstream of 3'-UTR).

Chr	SNP	Position (hg19)	P (PC1)	P (IQadjPC1)	Gene (distance)[a]
19	rs8106213	20657781	0.002	0.004	ZNF737(+63.01)
19	rs11669293	20663314	0.015	0.023	ZNF737(+57.48)
19	rs2021399	20682055	9×10^{-4}	3×10^{-4}	ZNF737(+38.74)
19	rs2545918	20691114	5×10^{-4}	9×10^{-4}	ZNF737(+29.68)
19	rs4809060	20701612	0.005	0.01	ZNF737(+19.19)
19	rs2545931	20704619	0.006	0.005	ZNF737(+16.18)
19	rs4809062	20707568	0.028	0.015	ZNF737(+13.23)
19	rs33948	20715228	0.007	0.006	ZNF737(+5.57)

Table S3e. Set of consecutive probes on 19p12 associated with PC1 and IQ-adjusted PC1 in the FamCNV analysis of the CLDRC-RD subset.
[a] Physical distance (kb) from *ZNF737* is indicated, along with orientation based on the direction of transcription ("-" = upstream of 5'-UTR, "+" = downstream of 3'-UTR).

Chr	SNP	Position (hg19)	Beta (PC1)[a]	P-value (PC1)	Beta (IQadjPC1)[a]	P-value (IQadjPC1)	Gene (distance)[b]
19	rs12610629	20626179	0.38	0.014	0.35	0.013	ZNF737(+94.61)
19	rs10408291	20630360	0.37	0.012	0.34	0.011	ZNF737(+90.43)
19	rs7254186	20631948	0.37	0.014	0.34	0.022	ZNF737(+88.84)
19	rs10403597	20647550	0.37	0.019	0.34	0.015	ZNF737(+73.24)
19	rs7251145	20656048	0.37	0.011	0.34	0.018	ZNF737(+64.74)
19	rs8106213	20657781	0.37	0.02	0.34	0.013	ZNF737(+63.01)
19	rs11669293	20663314	0.37	0.01	0.34	0.013	ZNF737(+57.48)
19	rs2021399	20682055	0.37	0.011	0.34	0.016	ZNF737(+38.74)
19	rs2545918	20691114	0.37	0.013	0.34	0.016	ZNF737(+29.68)
19	rs4809060	20701612	0.37	0.011	0.34	0.017	ZNF737(+19.19)
19	rs2545931	20704619	0.37	0.014	0.34	0.014	ZNF737(+16.18)
19	rs4809062	20707568	0.37	0.017	0.34	0.015	ZNF737(+13.23)
19	rs33948	20715228	0.39	0.009	0.36	0.014	ZNF737(+5.57)

Table S3f. Set of consecutive probes on 19p12 associated with PC1 and IQ-adjusted PC1 in the PLINK QFAM analysis of the CLDRC-RD subset.
[a] Beta values are indicative of the direction of effect of the "CNV+" state (i.e. copy number other than 2) but are not adjusted for family-based structure of the dataset, as per PLINK QFAM output. [b] Physical distance (kb) from *ZNF737* is indicated, along with orientation based on the direction of transcription ("-" = upstream of 5'-UTR, "+" = downstream of 3'-UTR).

Candidate pathway	Pathway size (nr of genes)	Overlaps	Empirical P	Corrected P
axonal guidance[a]	89	3	0.951	0.997
neuronal migration[b]	64	1	1	1
steroids[c]	333	8	0.999	1

Table S3g. Results of the pathway-based (INRICH) analysis of 913 CNV calls detected in 67 RD cases in the CLDRC dataset. In this analysis, three global composite candidate pathways were tested, representing specific neurobiological hypotheses on the etiology of reading and language disabilities: axon guidance, neuronal migration and steroid sex hormone biology. [a] All the GO sets containing the term "axon guidance". [b] All the GO sets containing the term "neuron migration". [c] All the GO sets containing the terms "steroid", "androgen", "estrogen", "progesterone" and "testosterone".

Chapter 6:
Assessing two novel candidate genes associated with reading and language in relation to cerebral cortical structure

This chapter is based on:

Gialluisi, A., Guadalupe, T., Francks, C. & Fisher, S.E. Assessing two novel candidate genes associated with reading and language in relation to cerebral cortical structure. (*in prep*)

Abstract

Structural and functional brain measures are gaining increasing attention in the study of reading and language cognition, as demonstrated by several imaging genetic studies on candidate RD/SLI genes. In the present chapter, we performed an imaging genetic analysis of two genes that showed the most significant associations in our Genome-Wide Association Scan Meta-Analysis (GWASMA) of reading and language skills (Chapter 3), namely *FLNC* (7q32.1) and *RBFOX2* (22q12.3).

In an independent dataset of healthy adults, we tested SNP associations with grey matter surface area and thickness of five cortical regions implicated in reading and language: middle temporal gyrus (MTG); pars opercularis and pars triangularis in the inferior frontal gyrus (IFG-PO and IFG-PT); postcentral parietal gyrus (PPG) and superior temporal gyrus (STG). For these regions, we also tested association with two different measures of asymmetry, an Asymmetry Index (AI) and its absolute value (AAI).

Analysis of the two most significantly associated SNPs from the reading and language GWASMA -rs59197085 (*FLNC*) and rs5995177 (*RBFOX2*)- revealed a significant multivariate association of rs5995177 with cortical thickness. This was driven by associations with left PPG, right MTG, right IFG-PT and IFG-PO, and in the STG bilaterally. The minor allele (A) -associated with reduced reading-language performance in our GWASMA- showed a negative effect on grey matter thickness, suggesting a potential link between these traits.

Gene-wide analysis of all the SNPs annotated to *FLNC* and *RBFOX2* revealed a borderline significant association between rs141148871 in *RBFOX2* and AAI of cortical thickness in the STG. The minor allele was associated with increased structural lateralization of the STG.

These results are consistent with the hypothesis that *RBFOX2*, which encodes a neuronal regulator of alternative splicing and is a potential target of FOXP2, may play a role in the neurobiology of reading and language, through genetic effects on cortical thickness. Further association analyses on reading/language traits, ideally combined with analysis of structural and functional brain imaging data in a single cohort, will help elucidate a potential role of this and other susceptibility genes in reading and language cognition.

Introduction

Structural and functional brain measures are gaining increasing attention in the study of neurodevelopmental disorders, including RD (Reading Disability, or dyslexia) and SLI (Specific Language Impairment). These disorders have been associated with variation in several neuroimaging measures, including brain connectivity, and measures of grey/white matter from structural Magnetic Resonance Imaging (MRI). In addition, the activation of specific brain regions has been investigated in relation to RD and SLI, both during performance of reading/language tasks -through functional MRI (fMRI)- and during resting state -through resting-state functional MRI (rs-fMRI). These traits may indicate some of the underlying neurobiological phenomena involved in RD and SLI. In other words, they may represent appropriate endophenotypes (as explained in Chapter 1), providing powerful means for the investigation of RD and SLI etiology. This approach has already been considered for various neuropsychiatric disorders (Thompson et al., 2010).

A finding that has received support from various neuroimaging studies of RD and SLI is that impaired individuals show a reduced average lateralization of language functions in the brain, compared to controls (as reviewed in Eicher & Gruen, 2013; Bishop, 2013). However, it is not yet clear whether this reduced functional brain asymmetry is more a cause or a consequence of reading and language deficits (Bishop, 2013). Neuroimaging studies also reported structural differences in both white and grey matter architecture in dyslexic and language impaired individuals, compared to non-impaired subjects (reviewed in Eicher & Gruen, 2013). The most frequently reported anomalies affect two brain regions typically involved in receptive and expressive language skills, namely Broca's and Wernicke's areas (Kennison, 2013). The former roughly corresponds to pars opercularis (PO) and pars triangularis (PT) in the left inferior frontal gyrus (IFG), while the latter overlaps with the posterior part of the left superior temporal gyrus (STG) (see Figure 2a in Chapter 1).

Reduced leftward asymmetries in the posterior part of the STG (also known as planum temporale) have been often associated with dyslexia, although not always consistently. Altarelli and colleagues (2014) recently meta-analysed previous neuroimaging studies on this region, reporting an altered pattern of asymmetry of the planum temporale surface area in dyslexic boys only, with a greater proportion of rightward asymmetrical cases compared to controls. In addition, Dole and colleagues (2013) reported a significant correlation between white matter density asymmetry in STG of dyslexic subjects and improved performance in speech-in-noise perception ability, linked to phonological processing. The central part of STG

also showed evidence of activation during speech comprehension tasks, in line with previous studies reporting a bilateral involvement of STG in speech recognition (Roux et al., 2014). STG anomalies have also been detected in language impaired children, characterized by smaller white matter volumes in the left hemisphere (Jancke et al., 2007) and smaller grey matter volumes bilaterally (Badcock et al., 2012). In contrast with this, comparison of affected versus unaffected subjects in a multiplex family with Childhood Apraxia of Speech (CAS, see Chapter 4) revealed a bilateral increase in grey matter density of STG for affected individuals (Belton et al., 2003; Watkins et al., 2002).

Similarly to STG, reduced grey matter leftward asymmetry has been observed also in the middle temporal gyrus (MTG) of dyslexic subjects (Dole et al., 2013), while another study reported reduced grey matter volume in the right MTG (Brambati et al., 2004). White matter anomalies have also been detected in this region, in SLI children (Soriano-Mas et al., 2009).

Another cortical region with a prominent role in phonological processing is the posterior part of the inferior frontal gyrus, where Broca's area is located (Salo et al., 2013; Lu et al 2007). Reduced grey matter volumes in left IFG and decreased leftward asymmetry have been reported both in RD (Hoeft et al., 2007; Brambati et al., 2006) and in CAS (Belton et al., 2003). This latter finding is in contrast with a report of SLI children exhibiting larger grey matter volumes in the left IFG (Badcock et al., 2012).

In light of the convergent neuroimaging evidence implicating the same brain regions in both RD and SLI, a neuroimaging analysis of comorbid RD-SLI cases was recently run to replicate these findings (Girbau-Massana et al., 2014). Surprisingly, the authors observed reduced grey matter volumes in the right postcentral parietal gyrus (PPG) and in medial occipital gyri bilaterally. Reduced grey matter volumes in right PPG were also observed in a group of SLI-only cases (Girbau-Massana et al., 2014), while a bilateral reduction of PPG was reported in RD cases versus controls (Hoeft et al., 2007). Consistently, an fMRI study reported an atypical bilateral activation of the PPG in language impaired children presenting with CAS (Liégeois et al., 2003).

In this chapter, we performed an imaging genetic analysis of the two genes that showed the most significant associations in our GWASMA of reading and language skills (Chapter 3), namely *FLNC* (7q32.1) and *RBFOX2* (22q12.3). This was aimed at detecting potential effects of these genes on brain architecture, and at assessing their consistency with structural brain anomalies reported to associate with RD/SLI (see above). We used structural MRI data from

a dataset of ~1,300 healthy adults (mean age ~24; Franke et al., 2010), to analyse genetic association with grey matter measures of five cortical regions implicated in reading and language by previous literature. These regions included middle temporal gyrus (MTG); pars opercularis and pars triangularis in the inferior frontal gyrus (IFG-PO and IFG-PT); postcentral parietal gyrus (PPG) and superior temporal gyrus (STG). Both left and right cortical measures -namely surface area and thickness- were produced for these regions, through automated segmentation and quantification of regional grey matter (Fischl et al., 2004). We carried out multivariate association analysis with these correlated measures, in order to reduce multiple testing separately region-by-region, and to detect potentially pleiotropic genetic effects on the cortical language networks constituted by these brain regions, while allowing for genetic effect sizes to vary across regions. We also tested for association using asymmetry indexes (defined for each region as (L-R)/(L+R)) and the absolute values of the AIs (i.e. unsigned magnitudes). The latter traits allowed us to detect potential genetic associations with the degree of structural lateralization, be it leftward or rightward.

Subjects and Methods

Dataset

The Brain Imaging Genetics (BIG) study was initiated in 2007 and comprises healthy volunteer subjects, including many university students, who participate in studies at the Donders Centre for Cognitive Neuroimaging, Nijmegen, The Netherlands (Franke et al., 2010). At the time of this study the BIG subject pool consisted of 2,337 self-reported healthy individuals (1,248 females) who had undergone anatomical (T1-weighted) MRI scans, usually as part of their involvement in diverse smaller scale studies at the Donders Center, and who had given their consent to participate in BIG. Their mean age at the time of first scan was 24.2 (SD 7.7; range 18-72). For the genetic analysis, genome-wide SNP genotype data were available from 1,276 of BIG subjects (see below for genotyping details). Their mean age was 22.9 years (SD 3.8; range 18-35), and 748 of these subjects were females. This dataset has already been used in other imaging genetics studies, investigating genetic associations with the asymmetry of planum temporale (Guadalupe et al., 2015) and of subcortical and hippocampal structures (Guadalupe et al., 2014b).

Phenotype elaboration and Quality Control (QC)

Image acquisition

MRI data were acquired in BIG as described elsewhere (Guadalupe et al., 2014a; 2014b; 2015). Briefly, MRI data acquisition was carried out with either a 1.5 Tesla Siemens Sonata or Avanto scanner or a 3 Tesla Siemens Trio or Tim Trio scanner (Siemens Medical Systems, Erlangen, Germany). Given that images were acquired during several smaller-scale studies, the parameters used were slight variations of a standard T1-weighted three-dimensional magnetization prepared rapid gradient echo sequence (MPRAGE; 1.0 x 1.0 x 1.0 mm voxel size). The most common variations in the TR/TI/TE/sagittal-slices parameters were the following: 2300/1100/3.03/192; 2730/1000/2.95/176; 2250/850/2.95/176; 2250/850/3.93/176; 2250/850/3.68/176; 2300/1100/3.03/192; 2300/1100/2.92/192; 2300/1100/2.96/192; 2300/1100/2.99/192; 1940/1100/3.93/176; and 1960/1100/4.58/176. There was also variation in the number of headcoils used across BIG scans, with the following arrays being employed (frequencies in brackets): 32-channel (26%), 12-channel (5%), 8-channel arrays (32%), and single headcoil (37%). For the genotyped sample, 634 subjects were scanned at 1.5 T, and 642 subjects at 3 T.

Image processing and phenotypic QC

Image processing has been described elsewhere (Guadalaupe et al., 2014a). Automated parcellation of cerebral cortical regions from T1-weighted images was done in FreeSurfer v5.3 (Fischl et al., 2004) according to the Desikan atlas (Desikan et al., 2006) within the "-recon-all" processing pipeline, and using default parameters. Measures of surface area (in mm^2) were produced for the total cortical surface and for each of 68 cortical parcellations, in each hemisphere. Regional measures of cortical thickness were also generated and analysed, as there is evidence that cortical surface and thickness have independent sources of variation (Panizzon et al., 2009). Estimates of Total Brain Volume (TBV) were calculated as the voxel-wise sum of the grey matter and white matter probability maps produced by the VBM8 toolbox, in SPM8 and with default settings. In line with previous imaging genetic association studies on this dataset (Guadalupe et al., 2014b; 2015), the following covariates were controlled for in subsequent analyses: gender, age, TBV, and field strength of the MRI (at either 1.5 or 3 T).

Cortical measures analysed

For the purposes of this chapter, we analysed both cortical thickness and surface area of the following regions: middle temporal gyrus (MTG); pars opercularis and pars triangularis in the inferior frontal gyrus (IFG-PO and IFG-PT); postcentral parietal gyrus (PPG) and superior temporal gyrus (STG), as defined in the Desikan atlas (Desikan et al., 2006). For each of these regions, we analysed left and right measures separately. These brain regions (highlighted in Figure 1) had been often reported to be involved in reading and language (dys)function in previous neuroimaging literature (see *Introduction* section). These measures showed moderate to high repeatability in scan-rescan correlation analysis of 342 twice-scanned subjects (0.62-0.76 and 0.84-0.91 for measures of cortical thickness and of cortical surface area, respectively) and generally moderate cross-trait correlations (see Table 1). Their distributions were approximately normal (absolute values of skewness and kurtosis <1 and <1.4, respectively) making them suitable for genetic association testing.

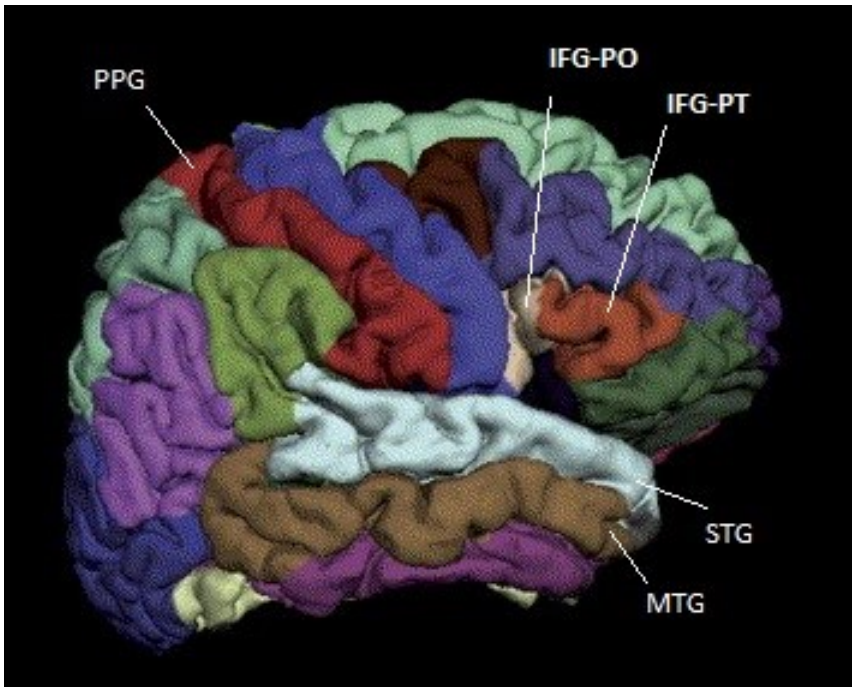


Figure 1. Cortical brain regions tested for association in the present chapter. Legend: MTG = middle temporal gyrus; IFG-PO = inferior frontal gyrus - pars opercularis; IFG-PT = inferior frontal gyrus - pars triangularis; PPG = postcentral parietal gyrus; STG = superior temporal gyrus.

Brain Measure[a]	MTG_ L	IFG-PO_ L	IFG-PT_ L	PPG_ L	STG_ L	MTG_ R	IFG-PO_ R	IFG-PT_ R	PPG_ R	STG_ R
MTG_L	1	0.143	0.069	0.225	0.26	0.577	0.124	0.076	0.232	0.355
IFG-PO_L	0.456	1	0.374	0.085	0.201	0.176	0.367	0.299	0.139	0.239
IFG-PT_L	0.419	0.554	1	0.045	0.186	0.031	0.212	0.437	0.077	0.187
PPG_L	0.357	0.367	0.378	1	0.241	0.26	0.111	0.053	0.478	0.254
STG_L	0.58	0.517	0.457	0.445	1	0.351	0.188	0.218	0.254	0.578
MTG_R	0.707	0.443	0.396	0.342	0.556	1	0.132	0.08	0.283	0.388
IFG-PO_R	0.398	0.551	0.404	0.338	0.491	0.432	1	0.212	0.094	0.206
IFG-PT_R	0.4	0.525	0.563	0.351	0.449	0.425	0.509	1	0.054	0.169
PPG_R	0.326	0.342	0.334	0.705	0.422	0.309	0.317	0.329	1	0.291
STG_R	0.561	0.531	0.465	0.47	0.758	0.615	0.506	0.47	0.438	1

Table 1. Cross-trait correlations of the brain measures tested, corrected for covariates used in the analysis (gender, age, TBV, and field strength of the MRI). The upper part of the matrix (above the diagonal) shows correlations across measures of cortical surface area, while the lower part (below the diagonal) refers to measures of cortical thickness.

[a]Legend: MTG = middle temporal gyrus; IFG-PO = inferior frontal gyrus - pars opercularis; IFG-PT = inferior frontal gyrus - pars triangularis; PPG = postcentral parietal gyrus; STG = superior temporal gyrus. Suffixes "L" and "R" indicate left and right hemisphere, respectively.

Genotype QC

Genotyping of BIG was performed as described in Guadalupe et al. (2014b; 2015). Briefly, genotype calls were generated using the Birdseed algorithm (Rabbee & Speed, 2006) on raw data from the Affymetrix Genome-Wide Human SNP Array 6.0 (Affymetrix Inc., Santa Clara, CA, USA). Samples were excluded that had call rates <90% and that showed deviant values of genome-wide heterozygosity. SNPs with a Minor Allele Frequency (MAF) <1% or that failed the Hardy-Weinberg equilibrium test (at a threshold $p \leq 10^{-6}$) were also excluded. The resulting markers were then adjusted to the forward strand, as to avoid any ambiguity problems in subsequent steps. A two-steps imputation protocol was followed. We used the software MACH for haplotype phasing and Minimac for the final imputation (Howie et al., 2012; Li et al., 2010), with the 1000 Genomes Phase 1 v3 EUR reference panel (The 1000 Genomes Consortium, 2012). All monomorphic markers were removed from the reference dataset. Individual genotype calls that had an imputation certainty <90% were removed, as were markers with an overall quality score (r^2) <0.3. As a final QC step, only markers with ≤5% missing data were selected. At the end of these procedures, genotypes were available for 1,276 subjects from BIG, for 6,131,824 SNPs spanning the genome. For the purpose of this study, we extracted all the SNPs falling within or close to *FLNC* and *RBFOX2*. To include potential regulatory regions in the analysis, also SNPs in the vicinity of these genes, up to 50

kb beyond the 5'- and 3'- Untranslated Regions (UTRs), were extracted. The final number of SNPs available for subsequent analyses was 177 for *FLNC* and 418 for *RBFOX2*.

Genetic association analyses

Association analysis with cortical surface area and thickness measures

We carried out multivariate genetic association tests using both left and right cortical thickness and surface area traits (see Table 2a, b and *Cortical measures analysed* paragraph) using TATES (*Trait-based Association Test that uses Extended Simes procedure*; Van der Sluis et al., 2013; http://ctglab.nl/software/tates). Thicknesses and areas were analyzed in separate multivariate tests. The TATES method is claimed to be optimal for detecting multivariate genetic associations affecting some, but not necessarily all, of a set of correlated phenotypes (Van der Sluis et al., 2013).

TATES combines the p-values obtained in univariate genetic association analysis on multiple (correlated) phenotypes, to produce one multivariate association p-value per SNP, while correcting for the correlations between the phenotypes. The univariate associations needed as input for TATES analysis were tested through *--linear* analysis in PLINK v1.07 (Purcell et al., 2007), controlling for the covariates age, gender, TBV and field strength of the MRI. This method regresses the phenotype score on the SNP genotype in an additive linear model, for each of the SNPs tested. Further details on this analysis can be found in PLINK documentation (http://pngu.mgh.harvard.edu/~purcell/plink/download.shtml).

We initially tested the top independent associations from our GWASMA of reading and language traits (Chapter 3), namely rs59197085 (7q32.1) and rs5995177 (22q12.3). Therefore we performed 4 separate tests as our primary hypotheses for this study, i.e. each of two SNPs in each of two multivariate association tests (for thicknesses and for areas). This resulted in a corrected α threshold of 0.0125.

Then, as an exploratory analysis, we carried out multivariate association analysis (TATES) for all 595 SNPs within *FLNC* and *RBFOX2* (i.e. including SNPs up to 50 kb beyond the 5'- and 3'-UTRs). To make an appropriate correction for multiple testing given the LD structure within each gene, we calculated the effective number of independent tests using the *Genetic Type I error calculator* (Li et al., 2012; http://statgenpro.psychiatry.hku.hk/gec/index.php), using our genotypes as input. The effective number of tests was determined as 67 (25 in

FLNC and 42 in *RBFOX2*), further multiplied by a factor of two for testing separately for thicknesses and areas as above. This resulted in a corrected α threshold of $3.7x10^{-4}$.

Association analysis with asymmetry measures

The multivariate analysis described above already allowed for genetic effects to be lateralized, i.e. to have quantitatively different effects on left and right measures. As a further analysis related to this, for each pair of bilateral measures, we first computed an Asymmetry Index (AI), as previously described in Guadalupe et al. (2015). This was calculated through the formula (L-R)/(L+R), where L and R were the left and right regional grey matter measure (either thickness or surface area) respectively. The values of AIs could range theoretically from -1 to +1, with negative values denoting a rightward asymmetry, positive values a leftward asymmetry and zero in the case of perfect symmetry. We also derived the absolute values (i.e. unsigned magnitudes) for each AI, which will be called Absolute Asymmetry Indexes (AAI) hereafter. The values of AAIs could range theoretically from 0 (i.e. perfect symmetry) to 1 (very pronounced leftward/rightward asymmetry). To make these traits suitable for genetic association analysis, they were further residualized against the covariates gender, age, TBV, and field strength and further rank-normalized through Blom's formula in SPSS® 20.0, to remove skewness and attain normality.

We tested for genetic association with AI and AAI of each cortical region separately rather than in a multivariate model (given low pairwise correlations for AIs and AAIs across cortical regions; see Tables S1a, b), first for our top GWASMA hits - rs59197085 and rs5995177- and then gene-wide in *FLNC* and *RBFOX2*. For the former analysis, we computed a significance threshold of $1.25x10^{-3}$, correcting for multiple testing of two SNPs and 20 cortical asymmetries in total (i.e. two asymmetry indexes for each of the cortical measures analysed, namely surface area and thickness, and for each of the 5 brain regions tested). For the gene-wide analysis, we used a corrected α threshold of $3.7x10^{-5}$ (taking into account 20 asymmetry traits and 67 independent SNPs in total, as computed above).

Results

SNP associations with cortical surface and thickness measures

We first analysed the two most significantly associated SNPs detected in our GWASMA of reading and language skills (Chapter 3), namely rs59197085 (7q32.1) and rs5995177 (22q12.3). Multivariate (and corresponding univariate) associations of these SNPs with surface area and thickness measures of the ten cortical regions -including both left and right middle temporal gyrus (MTG), pars opercularis (IFG-PO) and pars triangularis of the inferior frontal gyrus (IFG-PT), postcentral parietal gyrus (PPG) and superior temporal gyrus (STG)- are reported in Tables 2a, b. These revealed a significant multivariate association of rs5995177 ($p \sim 0.012$, A/G, minor allele A, MAF ~ 7.8%) with grey matter thickness, which survived correction for multiple testing of 2 SNPs and 2 independent multivariate association tests ($\alpha = 0.0125$, see above). This association was mainly driven by associations with left PPG, right MTG, right IFG (PO and PT), and STG bilaterally (see Table 2b). The minor allele (A) was associated with a reduction of grey matter thickness (see Table 2b).

After focusing on the top hits from our reading/language GWASMA, we extended our multivariate association analysis to all the 595 SNPs falling within or close to our candidate genes (up to 50 kb beyond the 5'- and 3'-UTRs), *FLNC* and *RBFOX2*. This gene-wide analysis did not reveal any significant association withstanding correction for multiple testing of 2 multivariate tests and a total of 67 independent SNPs tested in the two genes ($\alpha = 3.7 \times 10^{-4}$, see *Subjects and Methods* for details). The top associated SNPs are reported in Table S1c. The most significant multivariate associations ($p < 0.01$) were observed with cortical thickness for eight polymorphisms in *RBFOX2*, namely rs78563107, rs6000084, rs6000085, rs144006011, 22:36264632:D, rs77169229, rs149940336 and 22:36419124:D ($p = 4.3\text{-}8.2 \times 10^{-3}$). These SNPs were all in high LD among themselves ($r^2 > 0.8$) and in moderate LD with rs5995177 ($r^2 \sim 0.5$). No SNPs showed suggestive multivariate ($p < 0.01$) or univariate association ($p < 0.001$) with cortical surface areas. Similarly, we did not observe any suggestive association within *FLNC*, neither in the univariate nor in the multivariate tests.

2a)

Chr	SNP	Position	MAF (%)	Multivariate[a]	MTG_ L	IFG-PO_ L	IFG-PT_ L	PPG_ L	STG_ L	MTG_ R	IFG-PO_ R	IFG-PT_ R	PPG_ R	STG_ R
7	rs59197085	128460756	8.52	0.663 (NA)	0.729 (-8.15)	0.661 (8.14)	0.521 (-8.55)	0.542 (-18)	0.927 (-2.3)	0.103 (-38.42)	0.622 (-8.11)	0.928 (1.49)	0.921 (-2.83)	0.137 (33.53)
22	rs5995177	36309553	7.82	0.996 (NA)	0.904 (-2.83)	0.366 (-16.91)	0.404 (11.23)	0.174 (-40.3)	0.995 (0.14)	0.954 (1.37)	0.991 (0.18)	0.996 (0.09)	0.966 (1.21)	0.789 (-6.02)

2b)

Chr	SNP	Position	MAF (%)	Multivariate[a]	MTG_ L	IFG-PO_ L	IFG-PT_ L	PPG_ L	STG_ L	MTG_ R	IFG-PO_ R	IFG-PT_ R	PPG_ R	STG_ R
7	rs59197085	128460756	8.52	0.724 (NA)	0.466 (-0.009)	0.206 (-0.014)	0.664 (-0.005)	0.269 (-0.01)	0.44 (-0.009)	0.567 (-0.007)	0.379 (0.01)	0.603 (-0.006)	0.812 (-0.002)	0.658 (-0.005)
22	rs5995177	36309553	7.82	**0.012 (NA)**	0.143 (-0.019)	0.061 (-0.021)	0.117 (-0.019)	**0.021 (-0.021)**	**2.4×10^{-3} (-0.037)**	**0.049 (-0.025)**	**0.015 (-0.029)**	**9×10^{-3} (-0.032)**	0.313 (-0.01)	**2.3×10^{-3} (-0.038)**

Table 2. Multivariate and univariate associations of the two top hits from our reading/language GWASMA (Chapter 3), rs59197085 (7q32.1) and rs5995177 (22q12.3), with measures of cortical **a)** surface area and **b)** thickness of the brain regions tested. Association p-values are reported, with beta values of the minor allele (A for both SNPs) in brackets. Significant multivariate associations ($p < 0.0125$) and nominally significant univariate associations ($p < 0.05$) are highlighted in bold. [a] For multivariate association analysis, p-values as computed by the software TATES are reported, but no beta value was produced in the output.
Legend: MTG = middle temporal gyrus; IFG-PO = inferior frontal gyrus - pars opercularis; IFG-PT = inferior frontal gyrus - pars triangularis; PPG = postcentral parietal gyrus; STG = superior temporal gyrus. Suffixes "L" and "R" indicate left and right hemisphere, respectively.

SNP associations with cortical asymmetry measures

Again we first assessed our top reading/language GWASMA hits, rs59197085 and rs5995177, and later tested all the SNPs available within *FLNC* and *RBFOX2*.

Analysis of the top associated SNPs in our GWASMA (Table 3a, b) revealed two nominally significant associations at rs59197085 (A/G, minor allele A, MAF ~ 8.5%), one with classical asymmetry (AI) of cortical thickness in pars opercularis ($p \sim 0.038$) and one with absolute asymmetry (AAI) of cortical thickness in pars triangularis ($p \sim 0.038$). In both associations, the minor allele (A) showed a negative effect on the asymmetry indexes ($\beta = -0.004$ and -0.15, respectively). No nominally significant associations were observed at rs5995177 in this analysis, although association with AAI of cortical thickness in the middle temporal gyrus fell just short of nominal significance ($p \sim 0.067$). None of the associations mentioned above survived correction for multiple testing as computed above.

The most significant associations with brain asymmetries among all the SNPs annotated to *FLNC* and *RBFOX2* are reported in Tables 4 and 5, showing associations with Asymmetry Index (AI) and Absolute Asymmetry Index (AAI), respectively. No significant associations surviving correction for multiple testing of 20 asymmetry traits and 67 independent SNPs tested in the two genes (α = of 3.7×10^{-5}) were detected in the analysis of AI, neither with measures of cortical surface area nor with measures of cortical thickness (Table 4). In this analysis, the most significant association was observed between rs956119 (G/A, minor allele G, MAF ~ 8%, located in *RBFOX2*) and AI of cortical thickness in the postcentral parietal gyrus ($p = 1.2 \times 10^{-4}$). This SNP was located ~57 Kb far from rs5995177 and was in low LD with it ($r^2 = 0.11$), and the minor allele showed a positive effect on AI (see Table 4). However, we observed a borderline significant association with AAI of cortical thickness in the superior temporal gyrus, at rs141148871 ($p = 3.7 \times 10^{-5}$, C/T, minor allele C, MAF ~ 1%), within *RBFOX2*. The second most significant association, which fell just short of statistical significance, was observed with the same asymmetry measure at rs144606679 ($p = 3.8 \times 10^{-5}$, C/T, minor allele C, MAF ~ 1%). These SNPs were in perfect LD in *RBFOX2* ($r^2 = 1$), but were in low LD with rs5995177 ($r^2 = 0.17$ for both SNPs). The minor allele showed a positive effect on AAI for both SNPs.

3a)

SNP	rs59197085		rs5995177	
Brain Asymmetry[a]	Surface Area	Thickness	Surface Area	Thickness
MTG	0.274 (0.003)	0.866 (-2.7 x 10^{-4})	0.477 (-0.002)	0.761 (5 x 10^{-4})
IFG-PO	0.396 (0.005)	**0.038 (-0.004)**	0.483 (-0.004)	0.442 (0.002)
IFG-PT	0.46 (-0.004)	0.973 (7.4 x 10^{-5})	0.454 (0.004)	0.292 (0.002)
PPG	0.655 (-0.001)	0.23 (-0.002)	0.189 (-0.004)	0.119 (-0.003)
STG	0.151 (-0.004)	0.646 (-0.001)	0.771 (0.001)	0.915 (1.6 x 10^{-4})

3b)

SNP	rs59197085		rs5995177	
Brain Asymmetry[a]	Surface Area	Thickness	Surface Area	Thickness
MTG	0.64 (-0.034)	0.403 (0.061)	0.832 (0.016)	0.067 (0.135)
IFG-PO	0.708 (0.027)	0.615 (-0.036)	0.358 (-0.068)	0.932 (0.006)
IFG-PT	0.344 (0.069)	**0.038 (-0.15)**	0.599 (-0.039)	0.319 (0.074)
PPG	0.641 (-0.034)	0.154 (-0.104)	0.324 (-0.073)	0.913 (-0.008)
STG	0.638 (-0.034)	0.443 (0.056)	0.689 (0.03)	0.164 (0.103)

Table 3. Associations with brain asymmetries of the two top hits from our reading/language GWASMA (Chapter 3), rs59197085 (7q32.1) and rs5995177 (22q12.3). P-values of associations with **a)** Asymmetry Index (AI) and **b)** Absolute Asymmetry Index (AAI) are reported, with beta values of the minor allele (A for both SNPs) in brackets. Nominally significant associations ($p < 0.05$) are highlighted in bold. [a] Legend: MTG = middle temporal gyrus; IFG-PO = inferior frontal gyrus - pars opercularis; IFG-PT = inferior frontal gyrus - pars triangularis; PPG = postcentral parietal gyrus; STG = superior temporal gyrus.

Brain Asymmetry[a]	Chr	SNP	Position	A1	A2	P-value	Beta	MAF (%)
PPG	22	rs956119	36252267	g	A	1.2×10^{-4}	0.006	8.33
IFG-PO	7	rs1565629	128445370	a	G	3×10^{-4}	-0.006	14.06
IFG-PO	7	rs17165191	128451004	c	T	3.6×10^{-4}	-0.005	14.04
IFG-PO	7	rs17165198	128451741	c	T	3.6×10^{-4}	-0.005	14.04
IFG-PO	7	rs4487676	128469362	a	G	4.1×10^{-4}	-0.006	12.54
IFG-PO	7	rs3807132	128469484	c	T	4.1×10^{-4}	-0.006	12.54
IFG-PO	7	rs3807133	128469760	a	G	4.1×10^{-4}	-0.006	12.54
IFG-PO	7	rs2291573	128444820	g	A	4.2×10^{-4}	-0.006	13.76
IFG-PO	7	rs2307037	128449405	a	G	4.4×10^{-4}	-0.006	12.47
IFG-PO	7	rs4731515	128450592	c	T	4.4×10^{-4}	-0.006	12.47
IFG-PO	7	rs60389668	128452556	g	T	4.4×10^{-4}	-0.006	12.47
IFG-PO	7	rs2270593	128456234	t	C	4.4×10^{-4}	-0.006	12.47
IFG-PO	7	rs4731517	128458649	a	C	4.4×10^{-4}	-0.006	12.47
IFG-PO	7	rs56377531	128460745	t	C	4.4×10^{-4}	-0.006	12.47
IFG-PO	7	rs3734974	128461720	a	G	4.4×10^{-4}	-0.006	12.47
IFG-PO	7	rs730931	128462273	g	C	4.4×10^{-4}	-0.006	12.47
IFG-PO	7	rs58320939	128462843	a	T	4.4×10^{-4}	-0.006	12.47
IFG-PO	7	rs62479619	128463864	a	G	4.4×10^{-4}	-0.006	12.47
IFG-PO	7	rs62479620	128463971	t	C	4.4×10^{-4}	-0.006	12.47
IFG-PO	7	rs62479621	128465246	a	C	4.4×10^{-4}	-0.006	12.47
IFG-PO	7	rs60324735	128465450	g	A	4.4×10^{-4}	-0.006	12.47
IFG-PO	7	rs60894155	128465755	a	G	4.4×10^{-4}	-0.006	12.47
IFG-PO	7	rs4472439	128466635	a	G	4.4×10^{-4}	-0.006	12.47
IFG-PO	7	rs754920	128468412	a	G	4.4×10^{-4}	-0.006	12.47
IFG-PO	7	rs3823480	128468881	a	G	4.4×10^{-4}	-0.006	12.47
IFG-PO	7	rs17165226	128453626	g	C	4.4×10^{-4}	-0.006	12.48
IFG-PO	7	rs62479612	128454202	t	C	4.4×10^{-4}	-0.006	12.48
IFG-PO	7	rs2307036	128454537	a	C	4.4×10^{-4}	-0.006	12.48
IFG-PO	7	rs4728138	128468105	g	A	4.6×10^{-4}	-0.006	12.49

Table 4. Most significant associations with Asymmetry Index (AI) in the genes *FLNC* (7q32.1) and *RBFOX2* (22q12.3). Association p-values < 5×10^{-4} with AI of cortical thickness are reported, as no such associations were detected with AI of cortical surface area in any of the regions tested. Beta values refer to the minor allele (A1).

[a] Legend: MTG = middle temporal gyrus; IFG-PO = inferior frontal gyrus - pars opercularis; IFG-PT = inferior frontal gyrus - pars triangularis; PPG = postcentral parietal gyrus; STG = superior temporal gyrus.

Brain Asymmetry[a]	Chr	SNP[b]	Position	A1	A2	P-value	Beta	MAF (%)
STG	**22**	**rs141148871**	**36307890**	**c**	**t**	**3.7×10^{-5}**	**0.173**	**1.13**
STG	22	rs144606679	36336610	c	t	3.8×10^{-5}	0.165	1.13
STG	22	22:36201040:D	36201040	d	r	2×10^{-4}	0.164	1.29

Table 5. Most significant associations with Absolute Asymmetry Index (AAI) in the genes *FLNC* (7q32.1) and *RBFOX2* (22q12.3). Association p-values < 5×10^{-4} with AAI of cortical thickness are reported, as no such associations were detected with AAI of cortical surface area in any of the regions tested. Beta values refer to the minor allele (A1). Borderline significant association ($p \leq 3.7 \times 10^{-5}$) is highlighted in bold.
[a] Legend: MTG = middle temporal gyrus; IFG-PO = inferior frontal gyrus - pars opercularis; IFG-PT = inferior frontal gyrus - pars triangularis; PPG = postcentral parietal gyrus; STG = superior temporal gyrus. [b] Single-base indels were not filtered out from the imputed polymorphisms since they were reliably called in the imputation reference (1000 Genomes, Phase I v3), and were tested for association as they could represent coding frameshift variants of biological interest.

Discussion

In the present chapter, we analysed association of variants in the genes *FLNC* (filamin C, 7q32.1) and *RBFOX2* (RNA-binding protein, fox-1 homolog 2, 22q12.3) with structural brain measures. These genes had shown the strongest associations in our GWAS meta-analysis of reading and language traits (Chapter 3). Here we tested association with grey matter surface area and thickness of five cortical regions implicated in reading and language by previous neuroimaging literature, namely middle temporal gyrus (MTG); pars opercularis and pars triangularis in the inferior frontal gyrus (IFG-PO and IFG-PT); postcentral parietal gyrus (PPG) and superior temporal gyrus (STG).

A focused analysis of the top hits detected in our reading/language GWASMA (Chapter 3), namely rs59197085 (7q32.1, located ~10 kb upstream of *FLNC*) and rs5995177 (22q12.3, located within *RBFOX2*), revealed a significant multivariate association of rs5995177 with grey matter thickness. This suggested a generalized pleiotropic effect of rs5995177 on cortical thickness in the brain regions analysed, which was mainly driven by associations with left postcentral parietal gyrus, right middle temporal gyrus, right inferior frontal gyrus (both pars opercularis and pars triangularis), and in the superior temporal gyrus bilaterally. The minor allele (A) showed a negative effect on grey matter thickness. Interestingly, this was the same allele associated with reduced reading-language principal component score (PC1) in our GWASMA, which leads us to hypothesize a role of reduced cortical thickness in poor reading/language performance. The reduced cortical thickness found to be associated here with rs5995177 is consistent with the reduced grey matter volumes observed in reading and/or language impaired children for some of the regions tested in the present chapter,

including superior temporal gyrus (Badcock et al., 2012) and postcentral parietal gyrus (Girbau-Massana et al., 2014). The fact that rs5995177 was among the top associations with PC1 but not with IQ-adjusted PC1 in our GWASMA (see Chapter 3), allows us to hypothesize a potential influence of rs5995177 on cognitive domains underlying both language and general cognition -rather than on cognitive domains exclusively related to language- through its effect on cortical thickness. Other candidate SNPs in RD/SLI susceptibility genes have been reported to be associated with a generalized decrease of cognitive abilities and with reduced volumes in specific brain regions at the same time. Scerri and colleagues (2012) reported a significant association of the SNPs rs917235 and rs714939 in the *MRPL19/GCFC2* locus (2p12) with lower verbal IQ and with a bilateral decrease of white matter volume in the corpus callosum and in the cingulum. These SNPs had been previously associated with RD (Anthoni et al., 2007).

When we extended our analysis to all the SNPs annotated to *FLNC* and *RBFOX2*, association tests of measures of cortical surface area and thickness in the ten candidate brain regions revealed no significant associations surviving Bonferroni correction, neither at the univariate nor at the multivariate level.

Since poor reading and language performance has been associated with reduced lateralization of specific brain areas (Bishop, 2013), here we also wanted to test for potential effects of variants in *FLNC* and *RBFOX2* on the asymmetry of the brain regions tested. In addition to associations with a classical directional Asymmetry Index (AI) for each region, we also tested associations with its absolute value (AAI), so to detect potential genetic effects on the degree of lateralization, be it leftward or rightward. Association analysis of the two top hits of the GWASMA with AIs and AAIs of our candidate brain regions did not reveal any significant association surviving correction for multiple testing. However, rs59197085 showed two nominally significant associations in the inferior frontal gyrus, one with AI of cortical thickness in pars opercularis (IFG-PO) and one with AAI of cortical thickness in pars triangularis (IFG-PT). In both associations, the minor allele (A) -which had been associated to lower reading/language performance in our GWASMA (Chapter 3)- showed a negative effect on AI and AAI. In other words, it was associated with increased rightward lateralization in IFG-PO and with reduced absolute lateralization in IFG-PT. This, along with the nominal significance of these associations, suggests caution in the interpretation of these results, which may be due to type I error.

When we extended the analysis of asymmetries to all the SNPs in our two candidate genes, we observed a borderline significant association with AAI of cortical thickness in the superior temporal gyrus, at the intronic SNP rs141148871 in *RBFOX2*. Another close SNP, rs144606679, showed an association falling just short of statistical significance, tagging the same genetic effect as rs141148871 (see *Results* section). rs141148871 was in low LD with rs5995177, in spite of being quite close to it (~2 kb far). The minor allele (C) was associated with increased AAI, suggesting a positive effect on structural lateralization of the STG. The association was not replicated with AI, which is not surprising given the low pairwise correlation between AI and AAI in the STG (Pearson's r ~ -0.1). This SNP was not analysed in our previous reading/language GWASMA, and therefore we cannot make any resolute statement on its effect on reading and language skills.

Structural alterations in the language-related regions found to be associated in this chapter may reflect functional alterations, as changes in brain morphology have been often associated with experience-dependent plasticity in the Central Nervous System (CNS; Dole et al., 2013; Zatorre et al., 2012). Alternatively, such structural variations may be due to effects on neuronal migration (Eicher & Gruen, 2013). The fact that *RBFOX2* encodes an alternative splicing regulator very important in CNS development is consistent with a role of *RBFOX2* in reading and language cognition, via genetic effects on brain architecture. Interestingly, rs5995177 is located only ~1 kb far from a potential FOXP2 binding site (The ENCODE Project Consortium, 2012).

Additional brain regions -other than those tested here- have been implicated in reading and language, such as cerebellum (reviewed in Mariën et al., 2014), thalamus (reviewed in Klostermann et al., 2013), caudate nucleus (Vargha-Khadem et al., 1998; Watkins et al., 2002; Belton et al., 2003) and multiple fiber bundles which are thought to be important in creating a network among language-related areas of the brain (Vandermosten et al., 2012; Wandell & Yeatman, 2013; Girbau-Massana et al., 2014; Boets et al., 2013). Future imaging genetic analyses of *FLNC* and *RBFOX2* may include these regions, to test whether there are other genetic effects on brain measures which may be relevant to reading and language cognition.

More in general, further association analyses on reading and language traits in larger datasets, combined with the analysis of structural and functional brain imaging data in the same cohort, will help to elucidate the potential effects of *FLNC*, *RBFOX2* and other susceptibility genes on reading and language cognition, and to correlate these effects with changes in the architecture of brain regions underlying reading and language.

Acknowledgements

The authors declare no conflicts of interest. The BIG database was established in Nijmegen in 2007. This resource is now part of Cognomics, a joint initiative by researchers of the Donders Centre for Cognitive Neuroimaging, the Human Genetics and Cognitive Neuroscience departments of the Radboud University Medical Centre and the Max Planck Institute for Psycholinguistics. The Cognomics Initiative is supported by the participating departments and centres and by external grants, i.e. the Biobanking and Biomolecular Resources Research Infrastructure (Netherlands) (BBMRI-NL), the Hersenstichting Nederland and the Netherlands Organisation for Scientific Research (NWO). We also wish to thank all persons who kindly participated in this research.

Supplementary Material

- *S1: Supplementary Results.* Correlation matrices of asymmetry indexes (AIs and AAIs) of the brain regions tested. Top multivariate (and corresponding univariate) associations with cortical measures detected in the gene-wide analysis of *FLNC* and *RBFOX2*.

References

Altarelli, I., Leroy, F., Monzalvo, K., Fluss, J., Billard, C., Dehaene-Lambertz, G., Galaburda, A.M. & Ramus, F. (2014) Planum temporale asymmetry in developmental dyslexia: Revisiting an old question. *Hum Brain Mapp,* **35,** 5717-5735.

Anthoni, H., Zucchelli, M., Matsson, H., Muller-Myhsok, B., Fransson, I., Schumacher, J., Massinen, S., Onkamo, P., Warnke, A., Griesemann, H., Hoffmann, P., Nopola-Hemmi, J., Lyytinen, H., Schulte-Korne, G., Kere, J., Nothen, M.M. & Peyrard-Janvid, M. (2007) A locus on 2p12 containing the co-regulated MRPL19 and C2ORF3 genes is associated to dyslexia. *Hum Mol Genet,* **16,** 667-677.

Badcock, N.A., Bishop, D.V.M., Hardiman, M.J., Barry, J.G. & Watkins, K.E. (2012) Co-localisation of abnormal brain structure and function in specific language impairment. *Brain Lang,* **120,** 310-320.

Belton, E., Salmond, C., Watkins, K., Vargha-Khadem, F. & Gadian, D. (2003) Bilateral brain abnormalities associated with dominantly inherited verbal and orofacial dyspraxia. *Hum Brain Mapp,* **18,** 194 - 200.

Bishop, D.V. (2013) Cerebral asymmetry and language development: cause, correlate, or consequence? *Science,* **340,** 1230531.

Boets, B., Op de Beeck, H.P., Vandermosten, M., Scott, S.K., Gillebert, C.R., Mantini, D., Bulthé, J., Sunaert, S., Wouters, J. & Ghesquière, P. (2013) Intact But Less Accessible Phonetic Representations in Adults with Dyslexia. *Science,* **342,** 1251-1254.

Brambati, S.M., Termine, C., Ruffino, M., Danna, M., Lanzi, G., Stella, G., Cappa, S.F. & Perani, D. (2006) Neuropsychological deficits and neural dysfunction in familial dyslexia. *Brain Res,* **1113,** 174-185.

Brambati, S.M., Termine, C., Ruffino, M., Stella, G., Fazio, F., Cappa, S.F. & Perani, D. (2004) Regional reductions of gray matter volume in familial dyslexia. *Neurology,* **63,** 742-745.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S. & Killiany, R.J. (2006) An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, **31**, 968-980.

Dole, M., Meunier, F. & Hoen, M. (2013) Gray and White Matter Distribution in Dyslexia: A VBM Study of superior temporal gyrus Asymmetry. *PLoS ONE,* **8,** e76823.

Eicher, J.D. & Gruen, J.R. (2013) Imaging-genetics in dyslexia: Connecting risk genetic variants to brain neuroimaging and ultimately to reading impairments. *Mol Genet Metab,* **110,** 201-212.

Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D.H., Busa, E., Seidman, L.J., Goldstein, J., Kennedy, D., Caviness, V., Makris, N., Rosen, B. & Dale, A.M. (2004) Automatically Parcellating the Human Cerebral Cortex. *Cereb Cortex,* **14,** 11-22.

Franke, B., Vasquez, A.A., Veltman, J.A., Brunner, H.G., Rijpkema, M. & Fernández, G. (2010) Genetic Variation in CACNA1C, a Gene Associated with Bipolar Disorder, Influences Brainstem Rather than Gray Matter Volume in Healthy Individuals. *Biol Psychiatry,* **68,** 586-588.

Girbau-Massana, D., Garcia-Marti, G., Marti-Bonmati, L. & Schwartz, R.G. (2014) Gray–white matter and cerebrospinal fluid volume differences in children with Specific Language Impairment and/or Reading Disability. *Neuropsychologia,* **56,** 90-100.

Guadalupe, T., Willems, R.M., Zwiers, M.P., Arias Vasquez, A., Hoogman, M., Hagoort, P., Fernández, G., Buitelaar, J., Franke, B., Fisher, S.E. & Francks, C. (2014a) Differences in cerebral cortical anatomy of left- and right-handers. *Front Psychol,* **5**.

Guadalupe, T., Zwiers, M.P., Teumer, A., Wittfeld, K., Vasquez, A.A., Hoogman, M., Hagoort, P., Fernandez, G., Buitelaar, J., Hegenscheid, K., Völzke, H., Franke, B., Fisher, S.E., Grabe, H.J. & Francks, C. (2014b) Measurement and genetics of human subcortical and hippocampal asymmetries in large datasets. *Hum Brain Mapp,* **35,** 3277-3289.

Guadalupe, T., Zwiers, M.P., Wittfeld, K., Teumer, A., Vasquez, A.A., Hoogman, M., Hagoort, P., Fernandez, G., Buitelaar, J., van Bokhoven, H., Hegenscheid, K., Völzke, H., Franke, B., Fisher, S.E., Grabe, H.J. & Francks, C. (2015) Asymmetry within and around the human planum temporale is sexually dimorphic and influenced by genes involved in steroid hormone receptor activity. *Cortex,* **62,** 41-55.

Hoeft, F., Meyler, A., Hernandez, A., Juel, C., Taylor-Hill, H., Martindale, J.L., McMillon, G., Kolchugina, G., Black, J.M., Faizi, A., Deutsch, G.K., Siok, W.T., Reiss, A.L., Whitfield-Gabrieli, S. & Gabrieli, J.D.E.

(2007) Functional and morphometric brain dissociation between dyslexia and reading ability. *Proc Natl Acad Sci USA,* **104,** 4234-4239.

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet,* **44,** 955-959.

Jäncke, L., Siegenthaler, T., Preis, S. & Steinmetz, H. (2007) Decreased white-matter density in a left-sided fronto-temporal network in children with developmental language disorder: Evidence for anatomical anomalies in a motor-language network. *Brain Lang,* **102,** 91-98.

Kennison, S.M. (2013) Introduction to Language Development, SAGE Publications.

Klostermann, F. (2013) Functional Roles of the Thalamus for Language Capacities. *Front Syst Neurosci,* **7,** 32.

Li, M.X., Yeung, J.Y., Cherny, S. & Sham, P. (2012) Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum Genet*, **131**, 747-756.

Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol,* **34,** 816-834.

Liegeois, F., Baldeweg, T., Connelly, A., Gadian, D., Mishkin, M. & Vargha-Khadem, F. (2003) Language fMRI abnormalities associated with FOXP2 gene mutation. *Nat Neurosci*, **6**, 1230 - 1237.

Lu, L., Leonard, C., Thompson, P., Kan, E., Jolley, J., Welcome, S., Toga, A. & Sowell, E. (2007) Normal Developmental Changes in Inferior Frontal Gray Matter Are Associated with Improvement in Phonological Processing: A Longitudinal MRI Analysis. *Cereb Cortex,* **17,** 1092-1099.

Mariën, P., Ackermann, H., Adamaszek, M., Barwood, C.S., Beaton, A., Desmond, J., De Witte, E., Fawcett, A., Hertrich, I., Küper, M., Leggio, M., Marvel, C., Molinari, M., Murdoch, B., Nicolson, R., Schmahmann, J., Stoodley, C., Thürling, M., Timmann, D., Wouters, E. & Ziegler, W. (2014) Consensus Paper: Language and the Cerebellum: an Ongoing Enigma. *Cerebellum,* **13,** 386-410.

Panizzon, M.S., Fennema-Notestine, C., Eyler, L.T., Jernigan, T.L., Prom-Wormley, E., Neale, M., Jacobson, K., Lyons, M.J., Grant, M.D., Franz, C.E., Xian, H., Tsuang, M., Fischl, B., Seidman, L., Dale, A. & Kremen, W.S. (2009) Distinct Genetic Influences on Cortical Surface Area and Cortical Thickness. *Cereb Cortex,* **19,** 2728-2735.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. & Sham, P.C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet,* **81,** 559-575.

Roux, F.-E., Durand, J.-B., Réhault, E., Planton, S., Draper, L. & Démonet, J.-F. (2014) The neural basis for writing from dictation in the temporoparietal cortex. *Cortex,* **50,** 64-75.

Salo, E., Rinne, T., Salonen, O. & Alho, K. (2013) Brain activity during auditory and visual phonological, spatial and simple discrimination tasks. *Brain Res,* **1496,** 55-69.

Scerri, T.S., Darki, F., Newbury, D.F., Whitehouse, A.J.O., Peyrard-Janvid, M., Matsson, H., Ang, Q.W., Pennell, C.E., Ring, S., Stein, J., Morris, A.P., Monaco, A.P., Kere, J., Talcott, J.B., Klingberg, T. & Paracchini, S. (2012) The Dyslexia Candidate Locus on 2p12 Is Associated with General Cognitive Ability and White Matter Structure. *PLoS ONE,* **7,** e50321.

Soriano-Mas, C., Pujol, J., Ortiz, H., Deus, J., López-Sala, A. & Sans, A. (2009) Age-related brain structural alterations in children with specific language impairment. *Hum Brain Mapp,* **30,** 1626-1636.

The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.

The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature,* **491,** 56-65.

Thompson, P.M., Martin, N.G. & Wright, M.J. (2010) Imaging genomics. *Curr Opin Neurol* **23,** 368-373.

Van der Sluis, S., Posthuma, D. & Dolan, C.V. (2013) TATES: Efficient Multivariate Genotype-Phenotype Analysis for Genome-Wide Association Studies. *PLoS Genet,* **9,** e1003235.

Vandermosten, M., Boets, B., Wouters, J. & Ghesquière, P. (2012) A qualitative and quantitative review of diffusion tensor imaging studies in reading and dyslexia. *Neurosci Biobehav Rev,* **36,** 1532-1552.

Vargha-Khadem, F., Watkins, K.E., Price, C.J., Ashburner, J., Alcock, K.J., Connelly, A., Frackowiak, R.S.J., Friston, K.J., Pembrey, M.E., Mishkin, M., Gadian, D.G. & Passingham, R.E. (1998) Neural basis of an inherited speech and language disorder. *Proc Natl Acad Sci USA,* **95,** 12695-12700.

Wandell, B.A. & Yeatman, J.D. (2013) Biological development of reading circuits. *Curr Opin Neurobiol,* **23,** 261-268.

Watkins, K., Dronkers, N. & Vargha-Khadem, F. (2002) Behavioural analysis of an inherited speech and language disorder: comparison with acquired aphasia. *Brain,* **125,** 452 - 464.

Zatorre, R.J., Fields, R.D. & Johansen-Berg, H. (2012) Plasticity in gray and white: neuroimaging changes in brain structure during learning. *Nat Neurosci,* **15,** 528-536.

S1: Supplementary Results

S1a)

Brain Asymmetry[a]	MTG_AI	IFG-PO_AI	IFG-PT_AI	PPG_AI	STG_AI
MTG_AI	1	-0.01	0.041	0.023	-0.078
IFG-PO_AI	0.067	1	0.072	-0.063	-0.022
IFG-PT_AI	0.067	0.15	1	-0.036	-0.049
PPG_AI	-0.006	0.003	0.03	1	0.015
STG_AI	0.151	0.007	0.018	-0.02	1

S1b)

Brain Asymmetry[a]	MTG_AAI	IFG-PO_AAI	IFG-PT_AAI	PPG_AAI	STG_AAI
MTG_AAI	1	0.012	0.025	0.02	0.053
IFG-PO_AAI	0.001	1	-0.047	-0.008	-0.002
IFG-PT_AAI	0.042	0.036	1	0.065	0.002
PPG_AAI	0.01	0.042	-0.023	1	0.03
STG_AAI	0.015	0.049	0.036	0.045	1

Table S1. Cross-trait correlations of **a)** Asymmetry Indexes (AIs) and **b)** Absolute Asymmetry Indexes (AAIs) in the brain regions tested, corrected for covariates used in the analysis (gender, age, TBV, and field strength of the MRI). The upper part of each matrix (above the diagonal) shows correlations across measures of cortical surface area, while the lower part (below the diagonal) refers to measures of cortical thickness.
[a]Legend: MTG = middle temporal gyrus; IFG-PO = inferior frontal gyrus - pars opercularis; IFG-PT = inferior frontal gyrus - pars triangularis; PPG = postcentral parietal gyrus; STG = superior temporal gyrus.

S1c)

Chr	SNP[a]	Position	A1	A2	MAF (%)	Multivariate[b]	MTG_ L	IFG-PO_ L	IFG-PT_ L	PPG_ L	STG_ L	MTG_ R	IFG-PO_ R	IFG-PT_ R	PPG_ R	STG_ R
22	rs78563107	36449008	a	g	4.16	4.3×10^{-3}		0.028 (-0.032)			0.022 (-0.037)		7.5×10^{-3} (-0.041)			5.1×10^{-4} (-0.056)
22	rs6000084	36443943	t	a	4.26	7.1×10^{-3}		0.041 (-0.029)			0.045 (-0.032)		0.021 (-0.036)			8.5×10^{-4} (-0.053)
22	rs6000085	36444188	c	t	4.26	7.1×10^{-3}		0.041 (-0.029)			0.045 (-0.032)		0.021 (-0.036)			8.5×10^{-4} (-0.053)
22	rs144006011	36444625	t	c	4.26	7.1×10^{-3}		0.041 (-0.029)			0.045 (-0.032)		0.021 (-0.036)			8.5×10^{-4} (-0.053)
22	22:36264632:D	36264632	d	r	4.21	7.2×10^{-3}					0.029 (-0.035)		0.027 (-0.034)			8.7×10^{-4} (-0.053)
22	rs77169229	36269551	a	g	4.21	7.2×10^{-3}					0.029 (-0.035)		0.027 (-0.034)			8.7×10^{-4} (-0.053)
22	rs149940336	36449619	a	t	3.95	8.1×10^{-3}		0.033 (-0.032)			0.026 (-0.037)		9.2×10^{-3} (-0.041)			9.7×10^{-4} (-0.054)
22	22:36419124:D	36419124	d	r	4.25	8.2×10^{-3}					0.035 (-0.034)		0.043 (-0.031)			9.8×10^{-4} (-0.053)
22	rs114750168	36443613	t	c	4.3	8.4×10^{-3}					0.047 (-0.032)		0.028 (-0.034)			1×10^{-3} (-0.052)
22	rs6000082	36443675	t	c	4.3	8.4×10^{-3}					0.047 (-0.032)		0.028 (-0.034)			1×10^{-3} (-0.052)
22	rs8138352	36423512	c	g	4.24	8.4×10^{-3}					0.036 (-0.033)		0.042 (-0.031)			1×10^{-3} (-0.053)
22	rs118155841	36430131	t	g	4.3	8.5×10^{-3}					0.048 (-0.031)		0.028 (-0.034)			1×10^{-3} (-0.052)
22	rs8139721	36432952	a	g	4.3	8.5×10^{-3}					0.048 (-0.031)		0.028 (-0.034)			1×10^{-3} (-0.052)
22	rs5995190	36433664	t	c	4.3	8.5×10^{-3}					0.048 (-0.031)		0.028 (-0.034)			1×10^{-3} (-0.052)
22	rs56407516	36433837	a	g	4.3	8.5×10^{-3}					0.048 (-0.031)		0.028 (-0.034)			1×10^{-3} (-0.052)
22	rs8140210	36441964	t	c	4.3	8.5×10^{-3}					0.048 (-0.031)		0.028 (-0.034)			1×10^{-3} (-0.052)
22	rs8140469	36442122	t	c	4.3	8.5×10^{-3}					0.048 (-0.031)		0.028 (-0.034)			1×10^{-3} (-0.052)

Chr	SNP[a]	Position	A1	A2	MAF (%)	Multivariate[b]	MTG_ L	IFG-PO_ L	IFG-PT_ L	PPG_ L	STG_ L	MTG_ R	IFG-PO_ R	IFG-PT_ R	PPG_ R	STG_ R
22	rs73415795	36424026	t	g	4.23	8.6×10^{-3}					0.036 (-0.033)		0.043 (-0.031)			1×10^{-3} (-0.052)
22	rs77241789	36424232	g	t	4.23	8.6×10^{-3}					0.036 (-0.033)		0.043 (-0.031)			1×10^{-3} (-0.052)
22	rs117732943	36372075	t	c	4.25	8.8×10^{-3}					0.035 (-0.034)		0.046 (-0.031)			1.1×10^{-3} (-0.052)
22	rs118119033	36312453	t	c	4.22	9.5×10^{-3}					0.025 (-0.036)		0.034 (-0.033)			1.1×10^{-3} (-0.053)
22	rs77220577	36316373	a	g	4.21	9.7×10^{-3}					0.034 (-0.034)		0.036 (-0.033)			1.2×10^{-3} (-0.052)

Table S1c. Top multivariate associations (and corresponding univariate associations) with structural measures of ten candidate brain regions, in the genes *FLNC* (7q32.1) and *RBFOX2* (22q12.3). SNPs with multivariate association p-values < 0.01 are reported, along with univariate association p-values < 0.05 and corresponding beta values in brackets (referring to the minor allele, A1). These associations all refer to measures of cortical thickness, as no measure of cortical surface area showed multivariate association $p < 0.01$. Legend: MTG = middle temporal gyrus; IFG-PO = inferior frontal gyrus - pars opercularis; IFG-PT = inferior frontal gyrus - pars triangularis; PPG = postcentral parietal gyrus; STG = superior temporal gyrus. Suffixes "L" and "R" indicate left and right hemisphere, respectively.

[a] Single-base indels were not filtered out from the imputed polymorphisms since they were reliably called in the imputation reference (1000 Genomes, Phase I v3), and were tested for association as they could represent coding frameshift variants of biological interest. [b] For multivariate association analysis, p-values as computed by the software TATES are reported, but no beta value was produced in the output.

Chapter 7:
Summary and General discussion

Summary

This thesis was aimed at clarifying the genetic underpinnings shared between reading and language abilities. The existence of common genetic influences on reading and language has been suggested by several heritability studies and is supported also by partial phenotypic, clinical and biological overlaps between Reading Disability (RD) and Specific Language Impairment (SLI) (see Chapter 1). I investigated these common genetic bases by exploring associations with a measure of phenotypic variance shared across diverse continuous reading- and language-related traits (see below).

In **Chapter 2**, I investigated the relationship between various reading and language traits, and the evidence supporting shared biological bases. This analysis involved three datasets -two from the United Kingdom and one from Colorado (US)- which comprised children with reading or language problems and their siblings. Since there were moderate/strong intercorrelations among the reading and language traits analysed, I computed a first principal component score within each dataset (PC1), representing a notable proportion (52-75%) of the phenotypic variance shared across these traits. To have a measure of common variance independent of general (nonverbal) intelligence, I also derived a version of PC1 adjusted for performance IQ (IQ-adjusted PC1). An exploratory investigation of PC1, aimed in part at assessing its suitability to genetic association analysis, revealed three main findings: i) PC1 was highly correlated with the principal component score derived only from word reading and spelling ($PC1_{read}$), which were the only two measures available in all of the datasets and provided the closest phenotype matching possible across datasets; ii) dropping one or more traits from the PC1 computation did not substantially affect the resulting PC1 scores; iii) PC1 showed moderate to high heritabilities in all the datasets (0.29-0.84), in line with previous heritability estimates on RD, SLI and continuous reading and language traits. IQ-adjusted PC1 showed similar characteristics. To sum up, these analyses suggested the presence of a substantial phenotypic variance shared between reading and language skills, which is partly shared with general cognitive abilities and moderately influenced by genetic factors. PC scores -which well represent this common variance- were highly comparable across heterogeneous datasets, robust and heritable, supporting their suitability to genetic analyses.

In **Chapter 3**, I carried out a Genome Wide Association Scan Meta-Analysis (GWASMA) of PC1 and IQ-adjusted PC1, in order to detect Single Nucleotide Polymorphisms (SNPs) with pleiotropic effects on reading and language traits. Recently, two GWAS studies have been published with a similar purpose, one testing associations with quantitative reading and language traits in two population based cohorts (Luciano et al. 2013) and the other one testing association with RD-SLI comorbid cases through a classical case-control design (Eicher et al., 2013). Our GWASMA was complementary to these studies, as it investigated continuous trait variance across a broad range of reading and language abilities, with enrichment for poor performance. This study -which involved ~1,900 participants and ~5,5 million polymorphisms- detected suggestive associations at the SNPs rs59197085 and rs5995177 (uncorrected p ~10^{-7}), located respectively at the *CCDC136/FLNC* and *RBFOX2* genes. Both these SNPs showed evidence for effects across multiple reading and language traits in multivariate and univariate association tests against the individual traits used to compute PC1. In line with SNP associations, both *CCDC136/FLNC* and *RBFOX2* were among the top associated loci in the following gene-based association analysis. *RBFOX2* (22q12.3) is an important regulator of alternative splicing in neurons, while *FLNC* (7q32.1) encodes a structural protein involved in cytoskeleton remodelling. The *CCDC136/FLNC* locus also showed association with a comparable reading/language measure in an independent sample of 6,434 participants from the general population (previously analysed by Luciano et al., 2013), although involving distinct alleles of the associated SNP. Finally, a pathway-based association analysis using the results of the GWASMA -testing three candidate gene sets representing axon guidance, neuron migration and steroid-related pathways- revealed no significant enrichment of association signals.

In **Chapter 4**, I assessed associations of candidate SNPs and genes consistently implicated in RD/SLI by previous literature, and further investigated their cross-phenotypic effects on several reading and language traits, as above. At the SNP level, I observed nominally significant associations with PC scores (p ~10^{-2}-10^{-4}) for rs2143340, rs3212236, rs9461045 and rs761100 in *KIAA0319* (6p22.3); and for rs16973771, rs2875891 and rs8045507 in *ATP2C2* (16q24.1). These associations showed directions of effect consistent with the original studies, and a broad pleiotropy across reading and language traits - associations had been previously described in smaller subsets of our current datasets, and therefore this should be treated as an expanded analysis, rather than independent replication. An additional SNP,

rs12495133 in *ROBO1* (3p12) -which was recently found to be associated with RD and had never been investigated in our datasets- was associated with PC scores in our GWASMA (p ~10^{-4}), providing independent statistical support to the original association (Tran et al., 2014). Also this SNP showed cross-phenotypic effects on diverse reading and language traits and a concordant allelic trend with the original report. At the gene level, significant associations were detected for genes *KIAA0319* and *ROBO1*, in line with the results of the SNP-based assessment. Overall, these findings suggest pleiotropic effects of variants in *KIAA0319* and *ROBO1* across several reading and language traits, in line with previous studies. On the other hand, the majority of candidate SNPs tested did not show any evidence of association, in contrast with significant associations previously reported. This raises doubts on the replicability of the original findings, which will be discussed further below.

In **Chapter 5**, I investigated the genetic effects of Copy Number Variants (CNVs) on continuous reading and language traits, through a comprehensive analysis of the Colorado dataset (N~700), where CNVs were called using intensity data from ~723,000 DNA array probes. I first investigated correlations between CNV genomic burden and PC scores, which revealed no significant "global" influence of these variants on PC traits. Then I analysed associations with PC1/IQ-adjusted PC1 through two complementary genome-wide approaches. The first tested association between the CNV state at each probe and PC scores, considering both deletions and duplications at each location as a single CNV state, under the assumption that these would impact in the same way on reading/language performance. This analysis detected nominally significant associations (p ~10^{-2}-10^{-3}) within *CNTN4* (contactin 4) and *CTNNA3* (catenin alpha 3), which encode cell adhesion molecules with a likely role in neuronal development. An assessment of hotspot regions of CNVs involved in neuropsychiatric disorders (neuropsychiatric CNVs) allowed detection of an interval nominally associated with PC1 within *CHRNA7* (cholinergic nicotinic receptor alpha 7), encoding a ligand-gated ion channel mediating fast synaptic transmission. The second genome-wide analysis tested associations between raw intensity data for each probe and PC scores, to detect dosage-dependent effects of common multi-allelic CNVs in the genome. This revealed a region of association (p ~10^{-2}-10^{-4}) within a frequent deletion ~6 kb downstream of *ZNF737* (zinc finger protein 737, uncharacterized protein), which was also observed in the association analysis of CNV calls. Finally, a pathway-based association analysis of CNV calls detected in RD cases did not reveal any significant enrichment for the

three candidate gene sets previously tested in Chapter 3. Overall these data suggest that CNVs do not underlie a substantial proportion of variance in reading and language skills.

In **Chapter 6**, I carried out a follow-up imaging genetic analysis of the two genes showing the strongest associations in the GWASMA, namely *FLNC* and *RBFOX2*, in order to detect genetic effects of these genes on brain architecture. In an independent Dutch population-based cohort, I analysed both univariate and multivariate SNP associations with grey matter surface area and thickness of five cortical regions previously implicated in reading and language: middle temporal gyrus (MTG); pars opercularis and pars triangularis in the inferior frontal gyrus (IFG-PO and IFG-PT); postcentral parietal gyrus (PPG) and superior temporal gyrus (STG). For these regions, I also tested association with two different measures of asymmetry, an Asymmetry Index (AI, representing directional lateralization) and its absolute value (AAI, representing the degree of lateralization). One of the top hits in the GWASMA -rs5995177 (*RBFOX2*)- showed a significant multivariate association with cortical thickness, driven by univariate associations in left PPG, right MTG, right IFG-PT and IFG-PO, and in the STG bilaterally. The minor allele (A) -associated with reduced reading-language performance in our GWASMA- showed a negative effect on grey matter thickness across all the regions tested, suggesting a potential link between these traits. We also detected a borderline significant association with AAI of cortical thickness in the STG, at the SNPs rs141148871 (*RBFOX2*). These results suggest that *RBFOX2* may play a role in the neurobiology of reading and language, through genetic effects on cortical thickness.

General discussion

Novel candidate susceptibility loci

Although no genome-wide significant associations were detected in this thesis, novel candidate susceptibility genes with subtle effects on reading and language skills were found. *FLNC* and *RBFOX2* -identified in the GWASMA- show several direct and indirect links with reading and language cognition (see Chapter 3 for details). More prominently, *RBFOX2* (22q12.3) is a potential target of FOXP2 (The ENCODE Project Consortium, 2012), which is heavily implicated in speech and language disorders (Fisher & Scharff, 2009), and its homologue *RBFOX1* has been implicated in Rolandic Epilepsy, a typical neuronal migration disorder which is comorbid with RD and SLI (Clarke et al., 2007; Pal et al., 2011). Similarly,

FLNC (7q32.1) is the homologue of another gene, *FLNA*, previously implicated in periventricular heterotopia, a neuronal migration disorder sometimes associated with RD (Robertson, 2005; Poelmans et al., 2011). Moreover, independent studies have reported linkage of the 7q32 region with RD status (Kaminen et al., 2003) and with continuous reading-related traits, namely nonword spelling and irregular word reading (Bates et al., 2007). Interestingly, in Chapter 6 we also identified two independent genetic effects of *RBFOX2* on cortical thickness of candidate brain regions implicated in reading and language (see *Summary* above). One of them was detected at the local top hit in our GWASMA, rs5995177. This suggests that *RBFOX2*, encoding an alternative splicing regulator acting in neuronal development, may influence reading and language skills by affecting cortical thickness, a suggestion that can be followed up in larger neuroimaging genetics studies.

CNV analysis (Chapter 5) also provided novel candidate susceptibility genes with plausible roles in reading and language cognition, although again associations detected did not reach genome-wide significance. *CNTN4* (3p26.3) and *CTNNA3* (10q21.3) -showing some of the most significant associations in the genome-wide analysis of CNV state- have been already implicated in autism (ASD) and other neurodevelopmental disorders (Bacchelli et al., 2014; Glessner et al., 2009; Guo et al., 2012; Levy et al., 2011; Nava et al., 2014; Roohi et al., 2009; see Chapter 5 for details). Similarly, *CHRNA7* (15q13.3) -a hotspot region of neuropsychiatric CNVs which showed nominally significant association with CNV state- has been involved in various neurodevelopmental disorders, including ASD (Grayton et al., 2012; Griswold et al., 2012; Malhotra & Sebat, 2012), Attention Deficit Hyperactivity Disorder (ADHD; Williams et al., 2012) and epilepsy (Helbig et al., 2009). The above mentioned disorders all present some phenotypic, clinical or neurobiological overlap with RD and/or SLI (Willcutt et al. 2010; Mueller 2012; Bishop, 2010, Smith et al., 2012). This is consistent with the idea of partly shared etiologies across these disorders (reviewed in Chapter 1), and lends further support to the "generalist gene" hypothesis, maintaining that the same biological/genetic bases may subserve different cognitive functions through pleiotropic effects (Plomin & Kovas, 2005; Kovas & Plomin, 2006).

Contrasting results from the assessment of RD/SLI candidate genes

Although the assessment of candidate SNPs consistently implicated in RD/SLI revealed some significant associations with PC scores (in genes *KIAA0319*, *ATP2C2* and *ROBO1*), the

majority of candidate SNPs tested in Chapter 4 did not show any evidence of association. Even when significant associations were found, allelic trends were not always consistent with the original findings and across datasets (see Chapter 4 for further details). Similarly, previous genetic studies on reading and language phenotypes have reported scarce evidence of replication for the candidate SNPs tested in Chapter 4, both in genome-wide (Luciano et al., 2013; Eicher et al., 2013) and in targeted association analyses (Luciano et al. 2007; Paracchini et al., 2011; Becker et al., 2014; Tran et al., 2014). Some of these studies (reviewed in Chapter 4) have reported weak or no association and inconsistent allelic effects for our candidate SNPs, and the subsequent meta-analyses have led to partially inconclusive results (Zou et al., 2012; Tran et al., 2013; Zhong et al., 2013). While the original SNP associations were mainly reported in small samples (a few tens or hundreds of subjects), implying relatively large effect sizes, studies that failed to replicate these associations usually analysed samples between ~500 and ~6,000 subjects and had more power to detect such effects. All these elements put into question the replicability of the original findings.

Several reasons can be proposed to explain these inconsistencies across different studies.

A first plausible reason may be the heterogeneity of recruitment of samples analysed. The comparison between an RD/SLI case-control study and the analysis of continuous reading/language traits in a population based cohort may lead to discrepant results, as some genetic variants may have stronger effects in the lower tail of the distribution of these skills (i.e. in RD and SLI selected samples) and negligible effects in a broader range of variation (i.e. in general population samples), or vice versa. A second reason may be the heterogeneity of assessment of the traits analysed: different psychometric tests are often used in different studies to assess the same reading/language trait, which may lead to analyses of scores representing slightly discrepant cognitive abilities, and therefore introduce a bias in the comparison or meta-analysis of these studies. This applies also to the dichotomous classification of RD/SLI cases and controls, for which a consensus is far from being reached in the scientific community (Pennington & Bishop, 2009; Peterson & Pennington, 2012; Raskind et al., 2013). Third, different population genetic structure of the datasets analysed should be considered when comparing or meta-analyzing association studies. Variable LD patterns between the tag SNP (where the association is detected) and the genuine causal SNP (which determines the association) may occur in different populations, or even within the same (stratified) population, leading to contrasting allelic trends between studies (Lin et al., 2007; Luciano et al., 2007). Fourth, analyses of relatively small samples are more likely to

produce false positive findings (type I error; Colhoun et al., 2003). This may be a likely explanation especially for spurious associations, which have never been supported by further statistical (i.e. independent replications) or molecular evidence (i.e. functional validations).

A complex genetic background for complex disorders

Another apparently surprising result may be the lack of significant associations of candidate gene sets representing axon guidance, neuronal migration and sex hormones biology in pathway-based analyses. This was observed both in the enrichment test of association signals from the GWASMA (Chapter 3) and in the analysis of putative pathological CNVs detected in RD cases in the CNV study (Chapter 5). Of particular interest is the lack of association for axon guidance and neuron migration pathways, which have been linked with multiple candidate genes in RD and SLI (Hannula-Jouppi et al., 2005; Meng et al., 2005; Peschansky et al., 2010; Tammimies et al., 2013; Vernes et al., 2011; Poelmans et al., 2011; see Chapter 1 for an overview). These results suggest that SNPs and CNVs alone do not heavily affect these pathways. Nonetheless, we cannot rule out that these variants may still contribute to alter these molecular networks jointly with other kind of variants, such as rare mutations and Short Tandem Repeats (STRs), exerting very small individual effects.

The overall picture that we can draw from these findings is that common forms of RD and SLI have complex and partially overlapping genetic backgrounds, with hundreds or thousands of different genetic variants influencing reading and language abilities, usually through very small additive effects. To make this scenario complete, we should not forget the influence of environmental factors on these cognitive skills (see Chapter 1). Although the majority of these factors remains unknown, it is likely that they act both additively and jointly with genetic risk factors, through gene-by-environment (GxE) interactions, to affect reading and language performance (Pennington & Bishop, 2009). This picture (summarized in Figure 1) shall be taken into account in future research strategies aimed at clarifying the genetic bases of reading and language, as discussed below.

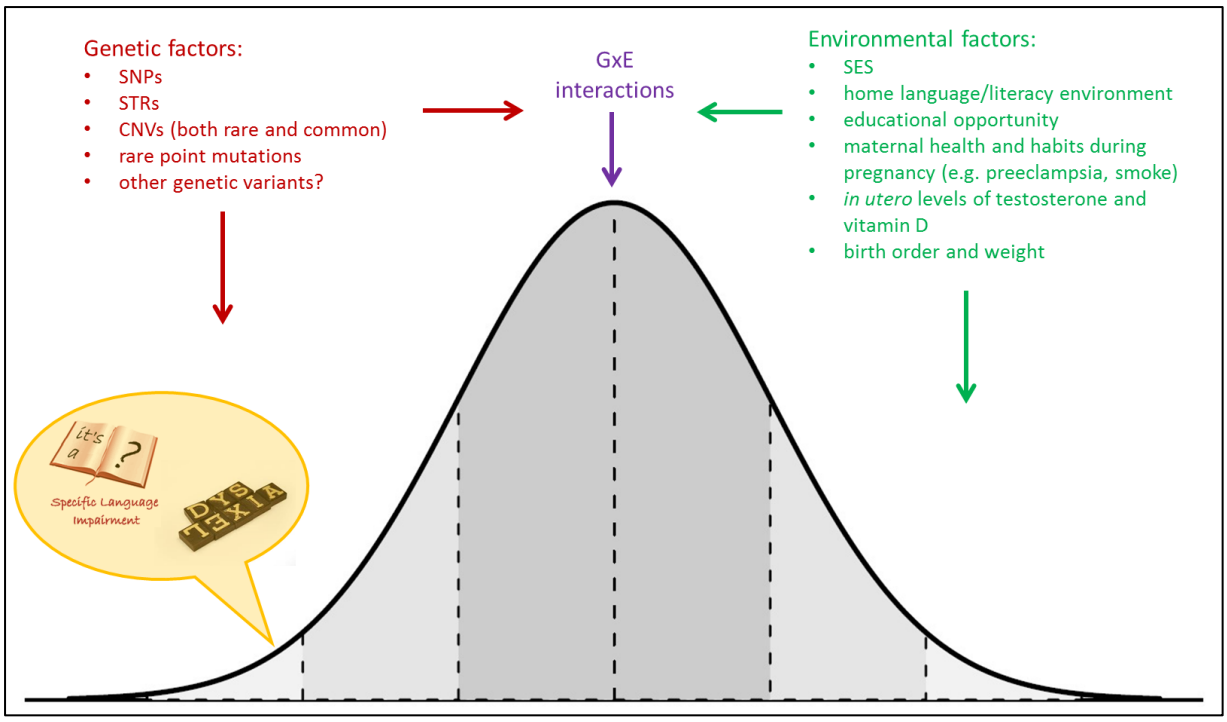


Figure 1. Abstract representation of the epidemiology of RD and SLI. Reading/language performance (and therefore liability to RD/SLI) is determined in an additive way by several risk and protective factors, which may be either genetic or environmental in origin. Genetic and environmental factors can also interact in a multiplicative way to increase or decrease reading/language performance, through GxE interactions.
Abbreviations: SNPs = Single Nucleotide Polymorphisms; STRs = Short Tandem Repeats; CNVs = Copy Number Variants; SES = socio-economic status; GxE = gene-by-environment interactions.

Future research strategies in the genetics of reading and language

In light of the elements mentioned above, I propose below some useful guidelines -some of which are already in use- for future investigations on the genetics of reading and language.

First, collecting large cohorts/datasets and characterizing them in detail at the phenotypic level is fundamental to investigate genetic associations with several cognitive skills underlying reading and language and to have enough power to detect even very small effects on these continuous traits. The availability of a wealth of reading and language traits, each representing distinct deficits which are thought to be at the basis of RD and SLI, may also allow testing of associations through a classical case-control design. This would improve the power to detect genetic variants which specifically affect the lower tail of the distribution of reading/language skills (i.e. categorical RD/SLI, see Figure 1), rather than continuous variation in the "normal performance" range.

Second, the use of comorbid cases may improve the power to detect variants affecting shared cognitive deficits between reading, language and other neurodevelopmental disorders. Eicher

et al. (2013) used this strategy in their GWAS on comorbid RD-SLI cases, but found no significant associations at the genome-wide level, probably due to the very low sample size of the study (less than 200 comorbid cases, see Chapter 3 for details). Another interesting strategy to identify genetic effects on cognitive deficits shared across disorders is the "mixed-GWAS" approach (Newbury et al., 2014), which analyzes different disorders in a single GWAS. A prominent example of this kind of study -analyzing ASD, ADHD, bipolar disorder, major depression and schizophrenia cases vs controls- led to the identification of putative overlapping genetic risk factors for several of these disorders (Cross-Disorder Group of the Psychiatric Genomic Consortium, 2013). My GWASMA strategy can be considered similar to this, although it focused on quantitative traits.

To increase the power to detect genetic variants with very small effect sizes (be they SNPs, CNVs or others), meta-analyzing different studies could represent an advantageous and cost-effective strategy. Nonetheless, this requires improving the homogeneity of studies, in terms of statistical methods used -both for association testing and for genotype quality control (QC)- and in terms of phenotypes analyzed. While statistical methods and QC are relatively easy to standardize, the heterogeneity of phenotypic assessment across different studies represents a notable hindrance for meta-analyses (as already discussed above), as this usually takes place a long time before the study and it is difficult and expensive to follow-up subjects through several sessions of assessment. In my GWASMA study, I obviated this issue by testing genetic associations with a principal component score representing common phenotypic variance in reading and language traits, which showed evidence of high robustness and comparability across heterogeneous datasets (Chapter 2, 3). More in general, trying to use homogeneous inclusion and diagnostic criteria in the studies, as well as universal psychometric tests to assess various cognitive traits, will surely make meta-analysis efforts more efficient in the future. Likewise, this may also help to shed a light on inconsistent results of candidate SNPs association studies.

An intrinsic limitation of association studies is that they are not informative about causality links between the associated variables. A procedure that may be useful in this direction is validating genetic associations through functional studies. These have already proven to be enlightening for diverse candidate SNPs implicated in RD and SLI, such as rs9461045 in *KIAA0319* (Dennis et al., 2009) and several SNPs in *ROBO1* (Hannula-Juppi et al., 2005), and may help to identify spurious associations, both for "old" and for "novel" candidate risk variants. Similarly, reporting and organizing even negative findings and inconsistent

associations of these variants into databases, as done by Bohland and colleagues (2014), will help to counterbalance the publication bias (i.e. significant results tend to be favored for publication) and the reporting bias (i.e. investigators tend to report only positive findings) in the field.

Finally, a "variant-wide" approach, which is more suited to the high genetic heterogeneity of RD and SLI, may allow researchers to clarify (at least in part) the missing heritability in reading and language skills. This implies investigating variants other than SNPs and CNVs, such as rare mutations, through high throughput sequencing technologies, in either multiplex families or sporadic cases. A prominent example of this approach is represented by a whole exome sequencing analysis of children affected by Childhood Apraxia of Speech (CAS), which reported suspected deleterious variants in several candidate susceptibility genes, including *CNTNAP2* and *KIAA0319* (Worthey et al., 2013). Further contributions to the understanding of risk factors influencing reading and language performance may come from investigating GxE interactions, as already done by Mascheretti et al. (2013).

Concluding remarks

In this thesis, I have brought to the attention of the reader several strategies aimed at clarifying the shared genetic underpinnings of reading and language, to get new insights in the genetic and neurobiological etiology of RD and SLI. Although no genetic associations were found that met genome-wide significance thresholds, I identified diverse novel candidate genes with subtle, pleiotropic effects on reading-language performance. One of these candidates (*RBFOX2*) also showed to affect cortical thickness, a plausible route through which it may influence reading and language abilities. Moreover, I made a contribution to clarifying the genetics of reading and language by assessing associations of several candidate SNPs previously implicated in RD/SLI, and by providing a detailed pattern of pleiotropy for those SNPs showing significant associations. Finally, I discussed various research strategies which can improve the power to detect genetic variants affecting written and oral language capacities. Overall, I believe that such a comprehensive approach will help to make progress in the understanding of language cognition and neurobiology.

References

Bacchelli, E., Ceroni, F., Pinto, D., Lomartire, S., Giannandrea, M., D'Adamo, P., Bonora, E., Parchi, P., Tancredi, R., Battaglia, A. & Maestrini, E. (2014) A CTNNA3 compound heterozygous deletion implicates a role for alphaT-catenin in susceptibility to autism spectrum disorder. *J Neurodevelop Disord*, **6**, 17.

Bates, T.C., Luciano, M., Castles, A., Coltheart, M., Wright, M.J. & Martin, N.G. (2007) Replication of reported linkages for dyslexia and spelling and suggestive evidence for novel regions on chromosomes 4 and 17. *Eur J Hum Genet,* **15,** 194-203.

Becker, J., Czamara, D., Scerri, T.S., Ramus, F., Csepe, V., Talcott, J.B., Stein, J., Morris, A., Ludwig, K.U., Hoffmann, P., Honbolygo, F., Toth, D., Fauchereau, F., Bogliotti, C., Iannuzzi, S., Chaix, Y., Valdois, S., Billard, C., George, F., Soares-Boucaud, I., Gerard, C.-L., van der Mark, S., Schulz, E., Vaessen, A., Maurer, U., Lohvansuu, K., Lyytinen, H., Zucchelli, M., Brandeis, D., Blomert, L., Leppanen, P.H.T., Bruder, J., Monaco, A.P., Muller-Myhsok, B., Kere, J., Landerl, K., Nothen, M.M., Schulte-Korne, G., Paracchini, S., Peyrard-Janvid, M. & Schumacher, J. (2014) Genetic analysis of dyslexia candidate genes in the European cross-linguistic NeuroDys cohort. *Eur J Hum Genet,* **22,** 675-680.

Bishop, D.V.M. (2010) Overlaps Between Autism and Language Impairment: Phenomimicry or Shared Etiology? *Behav Genet,* **40,** 618-629.

Bishop, D.V.M. & Snowling, M.J. (2004) Developmental Dyslexia and Specific Language Impairment: Same or Different? *Psychol Bull,* **130,** 858-886.

Bohland, J.W., Myers, E.M. & Kim, E. (2014) An Informatics Approach to Integrating Genetic and Neurological Data in Speech and Language Neuroscience. *Neuroinformatics,* **12,** 39-62.

Carrion-Castillo, A., Franke, B. & Fisher, S.E. (2013) Molecular Genetics of Dyslexia: An Overview. *Dyslexia*, **19**, 214-240.

Clarke, T., Strug, L.J., Murphy, P.L., Bali, B., Carvalho, J., Foster, S., Tremont, G., Gagnon, B.R., Dorta, N. & Pal, D.K. (2007) High Risk of Reading Disability and Speech Sound Disorder in Rolandic Epilepsy Families: Case–Control Study. *Epilepsia,* **48,** 2258-2265.

Colhoun, H.M., McKeigue, P.M. & Smith, G.D. (2003) Problems of reporting genetic associations with complex outcomes. *Lancet,* **361,** 865-872.

Cross-Disorder Group of the Psychiatric Genomic Consortium (2013) Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet,* **381,** 1371-1379.

Dennis, M.Y., Paracchini, S., Scerri, T.S., Prokunina-Olsson, L., Knight, J.C., Wade-Martins, R., Coggill, P., Beck, S., Green, E.D. & Monaco, A.P. (2009) A Common Variant Associated with Dyslexia Reduces Expression of the KIAA0319 Gene. *PLoS Genet,* **5,** e1000436.

Eicher, J.D., Powers, N.R., Miller, L.L., Akshoomoff, N., Amaral, D.G., Bloss, C.S., Libiger, O., Schork, N.J., Darst, B.F., Casey, B.J., Chang, L., Ernst, T., Frazier, J., Kaufmann, W.E., Keating, B., Kenet, T., Kennedy, D., Mostofsky, S., Murray, S.S., Sowell, E.R., Bartsch, H., Kuperman, J.M., Brown, T.T., Hagler, D.J., Dale, A.M., Jernigan, T.L., St. Pourcain, B., Davey Smith, G., Ring, S.M., Gruen, J.R., for the Pediatric Imaging, N. & Genetics, S. (2013) Genome-wide association study of shared components of reading disability and language impairment. *Genes Brain Behav,* **12,** 792-801.

Fisher, S. & Scharff, C. (2009) FOXP2 as a molecular window into speech and language. *Trends Genet,* **25,** 166 - 177.

Glessner, J.T., Wang, K., Cai, G., Korvatska, O., Kim, C.E., Wood, S., Zhang, H., Estes, A., Brune, C.W., Bradfield, J.P., Imielinski, M., Frackelton, E.C., Reichert, J., Crawford, E.L., Munson, J., Sleiman, P.M.A., Chiavacci, R., Annaiah, K., Thomas, K., Hou, C., Glaberson, W., Flory, J., Otieno, F., Garris, M., Soorya, L., Klei, L., Piven, J., Meyer, K.J., Anagnostou, E., Sakurai, T., Game, R.M., Rudd, D.S., Zurawiecki, D., McDougle, C.J., Davis, L.K., Miller, J., Posey, D.J., Michaels, S., Kolevzon, A., Silverman, J.M., Bernier, R., Levy, S.E., Schultz, R.T., Dawson, G., Owley, T., McMahon, W.M., Wassink, T.H., Sweeney, J.A., Nurnberger, J.I., Coon, H., Sutcliffe, J.S., Minshew, N.J., Grant, S.F.A., Bucan, M., Cook, E.H., Buxbaum, J.D., Devlin, B., Schellenberg, G.D. & Hakonarson, H. (2009) Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature*, **459**, 569-573.

Grayton, H.M., Fernandes, C., Rujescu, D. & Collier, D.A. (2012) Copy number variations in neurodevelopmental disorders. *Prog Neurobiol,* **99,** 81-91.

Griswold, A.J., Ma, D., Cukier, H.N., Nations, L.D., Schmidt, M.A., Chung, R.-H., Jaworski, J.M., Salyakina, D., Konidari, I., Whitehead, P.L., Wright, H.H., Abramson, R.K., Williams, S.M., Menon, R., Martin, E.R., Haines, J.L., Gilbert, J.R., Cuccaro, M.L. & Pericak-Vance, M.A. (2012) Evaluation of copy number variations reveals novel candidate genes in autism spectrum disorder-associated pathways. *Hum Mol Genet,* **21,** 3513-3523.

Guo, H., Xun, G., Peng, Y., Xiang, X., Xiong, Z., Zhang, L., He, Y., Xu, X., Liu, Y., Lu, L., Long, Z., Pan, Q., Hu, Z., Zhao, J. & Xia, K. (2012) Disruption of Contactin 4 in two subjects with autism in Chinese population. *Gene*, **505**, 201-205.

Hannula-Jouppi, K., Kaminen-Ahola, N., Taipale, M., Eklund, R., Nopola-Hemmi, J., Kääriäinen, H. & Kere, J. (2005) The Axon Guidance Receptor Gene *ROBO1* Is a Candidate Gene for Developmental Dyslexia. *PLoS Genet,* **1,** e50.

Harlaar, N., Hayiou-Thomas, M.E., Dale, P.S. & Plomin, R. (2008) Why Do Preschool Language Abilities Correlate With Later Reading? A Twin Study. *J Speech Lang Hear Res,* **51,** 688-705.

Helbig, I., Mefford, H.C., Sharp, A.J., Guipponi, M., Fichera, M., Franke, A., Muhle, H., de Kovel, C., Baker, C., von Spiczak, S., Kron, K.L., Steinich, I., Kleefuss-Lie, A.A., Leu, C., Gaus, V., Schmitz, B., Klein, K.M., Reif, P.S., Rosenow, F., Weber, Y., Lerche, H., Zimprich, F., Urak, L., Fuchs, K., Feucht, M., Genton, P., Thomas, P., Visscher, F., de Haan, G.J., Moller, R.S., Hjalgrim, H., Luciano, D., Wittig, M., Nothnagel, M., Elger, C.E., Nurnberg, P., Romano, C., Malafosse, A., Koeleman, B.P., Lindhout, D., Stephani, U., Schreiber, S., Eichler, E.E. & Sander, T. (2009) 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. *Nat Genet,* **41,** 160-162.

Kaminen, N., Hannula-Jouppi, K., Kestilä, M., Lahermo, P., Muller, K., Kaaranen, M., Myllyluoma, B., Voutilainen, A., Lyytinen, H., Nopola-Hemmi, J. & Kere, J. (2003) A genome scan for developmental dyslexia confirms linkage to chromosome 2p11 and suggests a new locus on 7q32. *J Med Genet,* **40,** 340-345.

Kovas, Y. & Plomin, R. (2006) Generalist genes: implications for the cognitive sciences. *Trends Cogn Sci,* **10,** 198-203.

Levy, D., Ronemus, M., Yamrom, B., Lee, Y., Leotta, A., Kendall, J., Marks, S., Lakshmi, B., Pai, D., Ye, K., Buja, A., Krieger, A., Yoon, S., Troge, J., Rodgers, L., Iossifov, I. & Wigler, M. (2011) Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron*, **70**, 886 - 897.

Lin, P.-I., Vance, J.M., Pericak-Vance, M.A. & Martin, E.R. (2007) No Gene Is an Island: The Flip-Flop Phenomenon. *Am J Hum Genet,* **80,** 531-538.

Logan, J., Petrill, S., Flax, J., Justice, L., Hou, L., Bassett, A., Tallal, P., Brzustowicz, L. & Bartlett, C. (2011) Genetic Covariation Underlying Reading, Language and Related Measures in a Sample Selected for Specific Language Impairment. *Behav Genet,* **41,** 651-659.

Luciano, M., Evans, D.M., Hansell, N.K., Medland, S.E., Montgomery, G.W., Martin, N.G., Wright, M.J. & Bates, T.C. (2013) A genome-wide association study for reading and language abilities in two population cohorts. *Genes Brain Behav,* **12,** 645-652.

Luciano, M., Lind, P.A., Duffy, D.L., Castles, A., Wright, M.J., Montgomery, G.W., Martin, N.G. & Bates, T.C. (2007) A haplotype spanning KIAA0319 and TTRAP is associated with normal variation in reading and spelling ability. *Biol Psychiatry,* **62,** 811-817.

Malhotra, D. & Sebat, J. (2012) CNVs: Harbingers of a Rare Variant Revolution in Psychiatric Genetics. *Cell,* **148,** 1223-1241.

Mascheretti, S., Bureau, A., Battaglia, M., Simone, D., Quadrelli, E., Croteau, J., Cellino, M.R., Giorda, R., Beri, S., Maziade, M. & Marino, C. (2013) An assessment of gene-by-environment interactions in developmental dyslexia-related phenotypes. *Genes Brain Behav,* **12,** 47-55.

Meng, H., Smith, S.D., Hager, K., Held, M., Liu, J., Olson, R.K., Pennington, B.F., DeFries, J.C., Gelernter, J., O'Reilly-Pol, T., Somlo, S., Skudlarski, P., Shaywitz, S.E., Shaywitz, B.A., Marchione, K., Wang, Y., Paramasivam, M., LoTurco, J.J., Page, G.P. & Gruen, J.R. (2005) DCDC2 is associated with reading disability and modulates neuronal development in the brain. *Proc Natl Acad Sci USA,* **102,** 17053-17058.

Mueller, K.L. (2012) Causation, correlation, or confound? What the comorbidity of language impairment and ADHD can tell us about the etiology of these disorders. University of Iowa.

Nava, C., Keren, B., Mignot, C., Rastetter, A., Chantot-Bastaraud, S., Faudet, A., Fonteneau, E., Amiet, C., Laurent, C., Jacquette, A., Whalen, S., Afenjar, A., Perisse, D., Doummar, D., Dorison, N., Leboyer, M., Siffroi, J.-P., Cohen, D., Brice, A., Heron, D. & Depienne, C. (2014) Prospective diagnostic analysis of copy number variants using SNP microarrays in individuals with autism spectrum disorders. *Eur J Hum Genet*, **22**, 71-78.

Newbury, D., Monaco, A. & Paracchini, S. (2014) Reading and Language Disorders: The Importance of Both Quantity and Quality. *Genes,* **5,** 285-309.

Pal, D.K. (2011) Epilepsy and neurodevelopmental disorders of language. *Curr Opin Neurol*, **24**, 126-131.

Paracchini, S., Ang, Q.W., Stanley, F.J., Monaco, A.P., Pennell, C.E. & Whitehouse, A.J. (2011) Analysis of dyslexia candidate genes in the Raine cohort representing the general Australian population. *Genes Brain Behav,* **10,** 158-165.

Pennington, B. & Bishop, D. (2009) Relations among speech, language and reading disorders. *Annu Rev Psych,* **60,** 283 - 306.

Peschansky, V.J., Burbridge, T.J., Volz, A.J., Fiondella, C., Wissner-Gross, Z., Galaburda, A.M., Turco, J.J.L. & Rosen, G.D. (2010) The Effect of Variation in Expression of the Candidate Dyslexia Susceptibility Gene Homolog Kiaa0319 on Neuronal Migration and Dendritic Morphology in the Rat. *Cereb Cortex,* **20,** 884-897.

Peterson, R.L. & Pennington, B.F. (2012) Developmental dyslexia. *Lancet,* **379,** 1997-2007.

Plomin, R. & Kovas, Y. (2005) Generalist Genes and Learning Disabilities. *Psychol Bull,* **131,** 592-617.

Poelmans, G., Buitelaar, J.K., Pauls, D.L. & Franke, B. (2011) A theoretical molecular network for dyslexia: integrating available genetic findings. *Mol Psychiatry,* **16,** 365-382.

Raskind, W.H., Peter, B., Richards, T.L., Eckert, M.A. & Berninger, V.W. (2013) The genetics of reading disabilities: from phenotypes to candidate genes. *Front Psychol,* **3,** 601.

Robertson, S.P. (2005) Filamin A: phenotypic diversity. *Curr Opin Genet Dev,* **15,** 301-307.

Roohi, J., Montagna, C., Tegay, D.H., Palmer, L.E., DeVincent, C., Pomeroy, J.C., Christian, S.L., Nowak, N. & Hatchwell, E. (2009) Disruption of contactin 4 in three subjects with autism spectrum disorder. *J Med Genet,* **46,** 176-182.

Smith, A.B., Kavros, P.M., Clarke, T., Dorta, N.J., Tremont, G. & Pal, D.K. (2012) A neurocognitive endophenotype associated with rolandic epilepsy. *Epilepsia,* **53,** 705-711.

Tammimies, K., Vitezic, M., Matsson, H., Le Guyader, S., Burglin, T.R., Ohman, T., Stromblad, S., Daub, C.O., Nyman, T.A., Kere, J. & Tapia-Paez, I. (2013) Molecular networks of DYX1C1 gene show connection to neuronal migration genes and cytoskeletal proteins. *Biol Psychiatry,* **73,** 583-590.

The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.

Tran, C., Gagnon, F., Wigg, K.G., Feng, Y., Gomez, L., Cate-Carter, T.D., Kerr, E.N., Field, L.L., Kaplan, B.J., Lovett, M.W. & Barr, C.L. (2013) A family-based association analysis and meta-analysis of the reading disabilities candidate gene DYX1C1. *Am J Med Genet B Neuropsychiatr Genet,* **162,** 146-156.

Tran, C., Wigg, K.G., Zhang, K., Cate-Carter, T.D., Kerr, E., Field, L.L., Kaplan, B.J., Lovett, M.W. & Barr, C.L. (2014) Association of the ROBO1 gene with reading disabilities in a family-based analysis. *Genes Brain Behav,* **13,** 430-438.

Vernes, S.C., Oliver, P.L., Spiteri, E., Lockstone, H.E., Puliyadi, R., Taylor, J.M., Ho, J., Mombereau, C., Brewer, A., Lowy, E., Nicod, J., Groszer, M., Baban, D., Sahgal, N., Cazier, J.-B., Ragoussis, J., Davies, K.E., Geschwind, D.H. & Fisher, S.E. (2011) Foxp2 Regulates Gene Networks Implicated in Neurite Outgrowth in the Developing Brain. *PLoS Genet,* **7,** e1002145.

Willcutt, E.G., Betjemann, R.S., McGrath, L.M., Chhabildas, N.A., Olson, R.K., DeFries, J.C. & Pennington, B.F. (2010) Etiology and neuropsychology of comorbidity between RD and ADHD: The case for multiple-deficit models. *Cortex,* **46,** 1345-1361.

Williams, N.M., Franke, B., Mick, E., Anney, R.J., Freitag, C.M., Gill, M., Thapar, A., O'Donovan, M.C., Owen, M.J., Holmans, P., Kent, L., Middleton, F., Zhang-James, Y., Liu, L., Meyer, J., Nguyen, T.T., Romanos, J., Romanos, M., Seitz, C., Renner, T.J., Walitza, S., Warnke, A., Palmason, H., Buitelaar,

J., Rommelse, N., Vasquez, A.A., Hawi, Z., Langley, K., Sergeant, J., Steinhausen, H.C., Roeyers, H., Biederman, J., Zaharieva, I., Hakonarson, H., Elia, J., Lionel, A.C., Crosbie, J., Marshall, C.R., Schachar, R., Scherer, S.W., Todorov, A., Smalley, S.L., Loo, S., Nelson, S., Shtir, C., Asherson, P., Reif, A., Lesch, K.P. & Faraone, S.V. (2012) Genome-wide analysis of copy number variants in attention deficit hyperactivity disorder: the role of rare variants and duplications at 15q13.3. *Am J Psychiatry,* **169,** 195-204.

Worthey, E., Raca, G., Laffin, J., Wilk, B., Harris, J., Jakielski, K., Dimmock, D., Strand, E. & Shriberg, L. (2013) Whole-exome sequencing supports genetic heterogeneity in childhood apraxia of speech. *J Neurodevelop Disord,* **5,** 29.

Zhong, R., Yang, B., Tang, H., Zou, L., Song, R., Zhu, L.-Q. & Miao, X. (2013) Meta-analysis of the Association Between DCDC2 Polymorphisms and Risk of Dyslexia. *Mol Neurobiol,* **47,** 435-442.

Zou, L., Chen, W., Shao, S., Sun, Z., Zhong, R., Shi, J., Miao, X. & Song, R. (2012) Genetic variant in KIAA0319, but not in DYX1C1, is associated with risk of dyslexia: An integrated meta-analysis. *Am J Med Genet B Neuropsychiatr Genet,* **159B,** 970-976.

Nederlandse samenvatting

Deze dissertatie heeft als doelstelling de genetische fundamenten, die gedeeld worden tussen lees- en taalvaardigheden, bloot te leggen. Het bestaan van zulke gemeenschappelijke genetische factoren in deze uniek menselijke eigenschap is eerder voorgesteld door verschillende erfelijkheidsonderzoeken en wordt ondersteund door de fenotypische, klinische en biologische overlap tussen Ernstige Leesstoornissen (EL; Reading Disability in het Engels) en Ernstige Spraak- en Taalmoeilijkheden (EST; Specific Language Impairment in het Engels) (zie Hoofdstuk 1).

In Hoofdstuk 2 heb ik de relatie tussen verschillende lees- en taaleigenvaardigheden in drie steekproeven onderzocht. Twee kwamen uit het Verenigd Koninkrijk en een derde uit Colorado (VS). Deze steekproeven bestonden uit kinderen met lees- en taalproblemen, samen met hun broers en/of zussen. De kinderen hun lees- en taalvaardigheden zijn eerst uitgebreid in beeld gebracht. Aan de hand hiervan heb ik binnen elke steekproef, door middel van statistische technieken, een primaire taalfactor afgeleid, welke de onderliggende lees- en taalvaardigheid representeert. Op eenzelfde wijze heb ik rekening gehouden met de verschillen in algemene intelligentie tussen de kinderen, om zo een maatstaf te formuleren die onafhankelijk van non-verbale cognitieve vaardigheden is. Ik heb de eigenschappen van deze factoren onderzocht om zodoende de vergelijkbaarheid van de steekproeven, de robuustheid van de analyse en de erfelijkheid van de lees- en taalvaardigheden in te schatten en heb hiermee de geschiktheid ervan voor genetische analyse aangetoond.

In Hoofdstuk 3 heb ik een genoombrede analyse op de taalfactoren in de eerdergenoemde steekproeven uitgevoerd. Dit soort analyses zijn bedoeld om verbanden te leggen tussen de verschillen in genetische achtergrond tussen personen en, in dit geval, de gemeten verschillen in lees- en taalvaardigheden. Hiervoor heb ik Enkel-Nucleotide Polymorfismen (ENPs; Single Nucleotide Polymosphisms in het Engels), d.w.z. een vaak voorkomende enkelvoudige verandering in een nucleotide (dit is het simpelste voorbeeld van genetische variatie) bekeken. Ik heb opmerkelijke associaties gevonden bij ENPs rs59197085 en rs5995177, respectievelijk gelokaliseerd in de CCDC136/FLNC en de RBFOX2 genen. Beide

ENPs voeren bewijs aan voor gemeenschappelijke effecten op lees- en taalvaardigheden. RBFOX2 (22q12.3) is een belangrijk regulatorgen met een speciale rol voor de ontwikkeling van neuronen, terwijl FLNC (7q32.1) betrokken is bij de vorming van de celstructuur. Beide genen tonen verscheidene directe en indirecte verbanden met lezen en taalcognitie (zie hieronder).

In Hoofdstuk 4 heb ik de associaties tussen kandidaat-ENPs en genen, d.w.z. de genen en ENPs die consequent door de literatuur over EL/EST worden aangemerkt, onderzocht om zodoende de consistentie met eerdere bevindingen te meten en om hun effecten op verscheidene lees- en taalvaardigheden te bepalen. Ik heb de kandidaat-ENP associaties ingeschat met behulp van de bovengenoemde taalfactoren door middel van meta-analyses (het samenvoegen van meerdere vergelijkbare analyses) op dezelfde steekproeven, en heb zodoende de pleiotropische (genen met meedere effecten op het fenotype) patronen van deze ENPs verder onderzocht, wat significante associaties aantoonde. Ik heb minder sterke associaties met de taalfactoren bij ENPs in de kandidaatgenen KIAA0319 (6p22.3), ROBO1 (3p12) en ATP2C2 (16q24.1) waargenomen. Deze ENPs lieten ook brede gemeenschappelijke effecten zien bij taal- en spraakvaardigheden, overeenkomend met eerdere verslagleggingen. Het meeste robuuste bewijs is gevonden bij de ENPs rs2143340 in KIAA0319 en rs12495133 in ROBO1, wat in de lijn der verwachtingen van eerdere lees- en taalsstudies ligt. Aan de andere kant toonde de meerderheid van de getoetste kandidaat-ENPs geen verband met lees- taalvaardigheid. Dit trekt de betrouwbaarheid van de eerdere bevindingen in twijfel.

In Hoofdstuk 5 heb ik de genetische effecten van copynumbervariaties (CNVs, een ander vorm van vaak voorkomende genetische variatie, d.w.z. kleine genomische series die een aantal keer worden herhaald bij verschillende mensen) bekeken. In de Colorado-steekproef heb ik eerst het verband tussen CNVs en taalfactoren onderzocht. In eerste instantie werd er geen globale invloed van de CNVs in het algemeen op taalfactoren gezien. Hierna heb ik de CNVs apart, en in meer detail, geanalyseerd. De resultaten uit deze analyse suggereren dat CNTN4 (contactine 4) en CTNNA3 (catenine alpha 3) waarschijnlijk een rol spelen in neuronale ontwikkeling. Hieropvolgende kandidaat-analyses, van CNVs betrokken bij

neuropsychiatische aandoeningen zoals autisme en aandachtsstoornisen, wijzen op een rol voor het CHRNA7 (cholinergishe nicotinische receptor alpha 7) gen, dat de overdracht van signalen tussen neuronen bemiddelt, en voor het ZNF737 (zinc vinger proteine 737, waarvan de functie nog niet bekend is) gen. Samengevat duiden deze subtiele bevindingen er op dat CNVs geen substantiele rol spelen bij lees- en taalvaardigheden.

In hoofdstuk 6 heb ik hersenscan-analyses uitgevoerd op de twee genen die de sterkste associaties toonden in de eerste analyses, namelijk FLNC and RBFOX2, om zo de mogelijke effecten van deze genen op de structuur van de hersenen te kunnen meten. Dit werd gedaan op een Nederlandse steekproef, bestaande uit gezonde, jonge volwassenen. In het bijzonder heb ik gekeken naar de hersengebieden die deel uitmaken van het taal-netwerk, namelijk, de gyrus temporalis medius, de pars opercularis, pars triangularis, de gyrus postcentralis (alleen het pariëtale gedeelte) en de gyrus temporalis superior. Deze gebieden heb ik geanalyseerd in termen van verschil in oppervlakte, dikte en asymmetrie tussen de linker en rechter hersenhelften. Het sterkste resultaat (alhoewel niet statistisch significant) heb ik gevonden bij rs5995177 (RBFOX2) in relatie met hersenschorsdikte van de linker gyrus temporalis medius, rechter pars opercularis, rechter pars triangularis en zowel de linker als rechter gyrus temporalis superior. Het gemeten effect van deze ENV (rs5995177) komt overeen met de bevindingen van eerdere analyses, namelijk dat een dunnere hersenschors in de genoemde gebieden samengaat met verminderde lees- en taalprestaties. Dit duidt op een mogelijk verband tussen deze eigenschappen.

Hoewel er geen statistisch significante resultaten zijn waargenomen in deze dissertatie, zijn er wel degelijk subtiele invloeden van de onderzochte genen op lees- en taalvaardigheden gevonden (Hoofdstuk 7). Ten eerste hebben FLNC and RBFOX2 mogelijk biologische en klinische verbanden met lezen en taalcognitie. Sterker nog, genen uit dezelfde familie als RBFOX2 en FLNC zijn betrokken bij psychiatrische stoornissen die veel kenmerken delen met EL en/of EST. Ten tweede hebben voorgaande onderzoeken ook een relatie aangetoond tussen regio 7q32 (waar FLNC is gelocaliseerd) en leesvaardigheid. Ten derde is RBFOX2 een mogelijk doel van FOXP2, een gen waarvan bekend is dat het een rol speelt in spraak- en taalstoornissen. Dit suggereert dat RBFOX2 betrokken zou kunnen zijn bij de biologische

fundamenten van lezen en taal middels een genetische invloed op hersenschorsdikte.

De algemene conclusie die we uit deze bevindingen kunnen trekken is dat typische vormen van EL en EST een complexe en gedeeltelijk overlappende genetische achtergrond hebben, waarbij honderden of duizenden verschillende genetische varianten hun invloed uitoefenen op lees- en taalvaardigheden. Dit gebeurt dan voornamelijk middels erg kleine opsommende effecten. Dit beeld zal in ogenschouw moeten worden genomen in toekomstige onderzoeksstrategieën gericht op het ophelderen van de genetische invloed op lezen en taalvaardigheid (Hoodstuk 7).

Curriculum Vitae

Alessandro Gialluisi was born in Castellana Grotte (Bari, Italy) on August 26th, 1984. In 2003, after finishing secondary school, he moved to Bologna to study biotechnology, and obtained a Bsc from the University of Bologna in 2007, with a research project on preimplantation genetic diagnosis in human ovocytes.

Then he moved to the University of Pisa, where he was awarded a Msc in Biomolecular science and technologies in 2010, with a final thesis on the genetic underpinnings of hypnotizability. In this period he also studied at the Complutense University of Madrid thanks to an Erasmus exchange fellowship, and did his final research internship at the Department of Physiological Sciences, University of Pisa.

After obtaining his Msc, he went back to Bologna, where he started his predoctoral research activity at the Medical Genetics Unit, Sant'Orsola-Malpighi Hospital, University of Bologna, in the field of medical genetics and genetic epidemiology of autosomal recessive disorders. In the same period, he worked as a private teacher for university courses and as a research assistant (junior data manager) at the Nephrology and Dialysis Unit, Malpighi Hospital, Bologna.

In September 2011 he started his PhD in Language and Genetics at the Max Planck Institute for Psycolinguistics, Nijmegen, under the supervision of Prof. Simon Fisher and Dr. Clyde Francks. Here he investigated the genetic basis of reading and language skills through statistical genetic studies, which represents the subject of this thesis. As a side-project, he showed interest in gene-culture coevolution, particularly in the persistence and transmission of recessive deafness and village sign languages in small and isolated communities.

He is currently a postdoctoral researcher in the Department of Translational Psychiatry at the Max Planck Institute of Psychiatry, Munich. Here, under the guidance of Prof Bertram Muller-Myhsok, he is continuing his research on the genetic basis of dyslexia and related cognitive skills through statistical genetic analyses.

Research contributions

Papers on international peer reviewed journals

- ✓ Alessandro Gialluisi, Dianne F. Newbury, Erik G. Wilcutt, Richard K. Olson, John C. DeFries, William M. Brandler, Bruce F. Pennington, Shelley D. Smith, Thomas S. Scerri, The SLI Consortium, Michelle Luciano, David M. Evans, Timothy C. Bates, John F. Stein, Joel B. Talcott, Anthony P. Monaco, Silvia Paracchini, Clyde Francks, Simon E. Fisher (2014). Genome-wide screening for DNA variants associated with reading and language traits. *Genes Brain and Behavior*, *13(7):686-701*.
- ✓ Tommaso Pippucci, Alberto Magi, Alessandro Gialluisi and Giovanni Romeo (2014). The analysis of WES data for the detection of ROH can be integrated into clinical and epidemiological studies based on consanguinity. *Human Heredity, 77(1-4):63-72*.
- ✓ Alberto Magi, Lorenzo Tattini, Flavia Palombo, Matteo Benelli, Alessandro Gialluisi, Betti Giusti, Rosanna Abbate, Marco Seri, Gian Franco Gensini, Giovanni Romeo and Tommaso Pippucci (2014). H3M2: Detection of Runs of Homozygosity from whole-exome sequencing data. *Bioinformatics, 30(20):2852-9.*
- ✓ Silvano Presciuttini, Alessandro Gialluisi, Serena Barbuti, Michele Curcio, Fabrizio Scatena, Giancarlo Carli and Enrica L. Santarcangelo (2014). Hypnotizability and Catechol-O-Methyltransferase (COMT) polymorphysms in Italians. *Frontiers in Human Neuroscience, 7: 929.*
- ✓ Gialluisi, A., Incollu, S., Pippucci, T., Lepori, M. B., Zappu, A., Loudianos, G., & Romeo, G. (2013). The homozygosity index (HI) approach reveals high allele frequency for Wilson disease in the Sardinian population. *European Journal of Human Genetics, 21(11):1308-11.*
- ✓ Gialluisi, A., Pippucci, T., & Romeo, G. (2013). Reply to ten Kate et al. *European Journal of Human Genetics*, *22(2):157-8*.
- ✓ Gialluisi, A., Dediu, D., Francks, C., & Fisher, S. E. (2013). Persistence and transmission of recessive deafness and sign language: New insights from village sign languages. *European Journal of Human Genetics, 21, 894-896.*

- ✓ Gialluisi, A., Pippucci, T., Anikster, Y., Ozbek, U., Medlej-Hashim, M., Mégarbané, A., & Romeo, G. (2012). Estimating the allele frequency of autosomal recessive disorders through mutational records and consanguinity: The homozygosity index (HI). *Annals of Human Genetics, 76, 159-167.*
- ✓ Romeo, G., Gialluisi, A., & Pippucci, T. (2012). Consanguinity studies and genome research in Mediterranean developing countries. *Middle East Journal of Medical Genetics, 1(1), 1-4.*

Presentations/posters at international conferences

- ✓ Gialluisi A., Newbury D.F., Wilcutt E.G., Olson R.K., Brandler W.M., Paracchini S., Monaco P., Francks C., Fisher S.E. "Genome-wide screening for DNA variants associated with dyslexia and language impairment", presentation at the 9th British Dyslexia Association International Conference, Guildford (UK), March 27-29, 2014 (http://bdainternationalconference.org/).
- ✓ A. Gialluisi, S. Incollu, T. Pippucci, M. B. Lepori, A. Zappu, G. Loudianos, G. Romeo. "The Homozygosity Index (HI) approach reveals high allele frequency for Wilson Disease in the Sardinian population", poster presented at the European Human Genetics Conference 2012, Nurberg (Germany) June 23-26, 2012. *European Journal of Human Genetics; June 2012, Vol 20, Suppl 1; 260.*
- ✓ G. Romeo, A. Gialluisi, T. Pippucci. "The past, present and future of consanguinity studies", presentation at The European Biotechnology Congress 2011, Instanbul (Turkey), Sept 28-Oct 1, 2011. *Current Opinion in Biotechnology; Sept 2011, Vol 22 (Suppl 1); S26.*
- ✓ Presciuttini S, Gialluisi A, Barbuti S, Curcio M, Scatena F, Ghelarducci B, Santarcangelo EL, Carli G. "Hypnotizability and polymorphisms of the COMT gene: an association study", poster presented at The 62nd National Congress of the Italian Physiological Society, Sorrento (Italy), Sept 25-27, 2011. *Acta Physiologica 2011; Volume 203, Supplement 688*.

- ✓ A. Gialluisi, T. Pippucci, Y. Anikster, U. Ozbek, M. Medlej-Hashim, A. Mégarbané, G. Romeo. "Estimating the prevalence of autosomal recessive disorders through mutational records and consanguinity: the Homozygosity Index (HI)", poster presented at the European Human Genetics Conference 2011, Amsterdam (Netherlands) May 28-31, 2011. *European Journal of Human Genetics; May 2011, Vol 19, Suppl 2; 336.*
- ✓ Graziano C, Gialluisi A, Baldazzi L, Balsamo A, Menabò S, Romeo G. "Estimating the prevalence of 21-hydroxylase deficiency in Italy through a novel molecular epidemiology approach", presentation at the 4th Yazd International Congress and Student Award in Reproductive Medicine, Yazd (Iran), April 15-17, 2011. *Iranian Journal of Reproductive Medicine, Spring 2011, Vol 9 (Suppl 2); 16-17.*
- ✓ G. Romeo, A. Gialluisi, T. Pippucci. "Consanguinity studies and genome research Mediterranean Developing Countries", presentation at the 1st International Conference of Human Genetics and Genome Research Division and National Society of Human Genetics "Human Genetics and Genomics: challenges for proper health", Cairo, November 8-10, 2010 (http://www.humangenetics2010.org/).
- ✓ T. Pippucci, A. Gialluisi, G. Romeo. "Estimating disease allele frequencies from mutational records in autosomal recessive disorders", presentation at the International Workshop on consanguinity: Health impact and future prospects for research, Geneva (Switzterland) May 3-7, 2010 (http://www.medecine.unige.ch/consanguinity/index.php).

Acknowledgements

It is really incredible how time is flying. Four years ago I arrived here as a young student, and now I am virtually leaving this place having reached a small but important milestone in my short career.

The first “thank you” goes to Simon and Clyde for making this possible, and for all the patience that they showed in these years...

And I also want to thank all the L&G and MPI people who shared with me even a single moment of this long walk, both in the most exciting and in the most depressing days…

A big thanks goes also to the Netherlands, a small but hospitable country where my daughter spent the first months of her life..for this reason, you will always have a place in my hearth..

And to my Italian friends here in Nijmegen..you made me feel home, and you know this is oxygen for an expat..

A sweet thank you to my high school crew in Apulia, to be always there, together, as we used to be 15 years ago..my soul is still with you and will always stay there..

A grateful thank you to my previous teachers and professors, for inspiring my passion for science..

And an endless thank you to my parents, for allowing me to receive an appropriate education, and to become an aware European citizen..

As I mentioned, four years ago I arrived here as a young student with no idea of his future, and now I leave this place as a father of two, and all of this thanks to you..

“tu mi hai aperto come una ferita, sto sanguinando, ma non ti lascio più..”

MPI Series in Psycholinguistics

1. The electrophysiology of speaking: Investigations on the time course of semantic, syntactic, and phonological processing. *Miranda van Turennout*
2. The role of the syllable in speech production: Evidence from lexical statistics, metalinguistics, masked priming, and electromagnetic midsagittal articulography. *Niels O. Schiller*
3. Lexical access in the production of ellipsis and pronouns. *Bernadette M. Schmitt*
4. The open-/closed-class distinction in spoken-word recognition. *Alette Haveman*
5. The acquisition of phonetic categories in young infants: A self-organising artificial neural network approach. *Kay Behnke*
6. Gesture and speech production. *Jan-Peter de Ruiter*
7. Comparative intonational phonology: English and German. *Esther Grabe*
8. Finiteness in adult and child German. *Ingeborg Lasser*
9. Language input for word discovery. *Joost van de Weijer*
10. Inherent complement verbs revisited: Towards an understanding of argument structure in Ewe. *James Essegbey*
11. Producing past and plural inflections. *Dirk Janssen*
12. Valence and transitivity in Saliba: An Oceanic language of Papua New Guinea. *Anna Margetts*
13. From speech to words. *Arie van der Lugt*
14. Simple and complex verbs in Jaminjung: A study of event categorisation in an Australian language. *Eva Schultze-Berndt*
15. Interpreting indefinites: An experimental study of children's language comprehension. *Irene Krämer*
16. Language-specific listening: The case of phonetic sequences. *Andrea Weber*
17. Moving eyes and naming objects. *Femke van der Meulen*
18. Analogy in morphology: The selection of linking elements in Dutch compounds. *Andrea Krott*
19. Morphology in speech comprehension. *Kerstin Mauth*
20. Morphological families in the mental lexicon. *Nivja H. de Jong*
21. Fixed expressions and the production of idioms. *Simone A. Sprenger*
22. The grammatical coding of postural semantics in Goemai (a West Chadic language of Nigeria). *Birgit Hellwig*

23. Paradigmatic structures in morphological processing: Computational and cross-linguistic experimental studies. *Fermín Moscoso del Prado Martín*
24. Contextual influences on spoken-word processing: An electrophysiological approach. *Daniëlle van den Brink*
25. Perceptual relevance of prevoicing in Dutch. *Petra M. van Alphen*
26. Syllables in speech production: Effects of syllable preparation and syllable frequency. *Joana Cholin*
27. Producing complex spoken numerals for time and space. *Marjolein Meeuwissen*
28. Morphology in auditory lexical processing: Sensitivity to fine phonetic detail and insensitivity to suffix reduction. *Rachèl J. J. K. Kemps*
29. At the same time...: The expression of simultaneity in learner varieties. *Barbara Schmiedtová*
30. A grammar of Jalonke argument structure. *Friederike Lüpke*
31. Agrammatic comprehension: An electrophysiological approach. *Marlies Wassenaar*
32. The structure and use of shape-based noun classes in Miraña (North West Amazon). *Frank Seifart*
33. Prosodically-conditioned detail in the recognition of spoken words. *Anne Pier Salverda*
34. Phonetic and lexical processing in a second language. *Mirjam Broersma*
35. Retrieving semantic and syntactic word properties. *Oliver Müller*
36. Lexically-guided perceptual learning in speech processing. *Frank Eisner*
37. Sensitivity to detailed acoustic information in word recognition. *Keren B. Shatzman*
38. The relationship between spoken word production and comprehension. *Rebecca Özdemir*
39. Disfluency: Interrupting speech and gesture. *Mandana Seyfeddinipur*
40. The acquisition of phonological structure: Distinguishing contrastive from non-contrastive variation. *Christiane Dietrich*
41. Cognitive cladistics and the relativity of spatial cognition. *Daniel B.M. Haun*
42. The acquisition of auditory categories. *Martijn Goudbeek*
43. Affix reduction in spoken Dutch. *Mark Pluymaekers*
44. Continuous-speech segmentation at the beginning of language acquisition: Electrophysiological evidence. *Valesca Kooijman*
45. Space and iconicity in German Sign Language (DGS). *Pamela Perniss*

46. On the production of morphologically complex words with special attention to effects of frequency. *Heidrun Bien*
47. Crosslinguistic influence in first and second languages: Convergence in speech and gesture. *Amanda Brown*
48. The acquisition of verb compounding in Mandarin Chinese. *Jidong Chen*
49. Phoneme inventories and patterns of speech sound perception. *Anita Wagner*
50. Lexical processing of morphologically complex words: An information-theoretical perspective. *Victor Kuperman*
51. A grammar of Savosavo, a Papuan language of the Solomon Islands. *Claudia Wegener*
52. Prosodic structure in speech production and perception. *Claudia Kuzla*
53. The acquisition of finiteness by Turkish learners of German and Turkish learners of French: Investigating knowledge of forms and functions in production and comprehension. *Sarah Schimke*
54. Studies on intonation and information structure in child and adult German. *Laura de Ruiter*
55. Processing the fine temporal structure of spoken words. *Eva Reinisch*
56. Semantics and (ir)regular inflection in morphological processing. *Wieke Tabak*
57. Processing strongly reduced forms in casual speech. *Susanne Brouwer*
58. Ambiguous pronoun resolution in L1 and L2 German and Dutch. *Miriam Ellert*
59. Lexical interactions in non-native speech comprehension: Evidence from electro-encephalography, eye-tracking, and functional magnetic resonance imaging. *Ian FitzPatrick*
60. Processing casual speech in native and non-native language. *Annelie Tuinman*
61. *Split intransitivity in Rotokas, a Papuan language of Bougainville. Stuart Robinson*
62. Evidentiality and intersubjectivity in Yurakaré: An interactional account. *Sonja Gipper*
63. The influence of information structure on language comprehension: A neurocognitive perspective. *Lin Wang*
64. The meaning and use of ideophones in Siwu. *Mark Dingemanse*
65. The role of acoustic detail and context in the comprehension of reduced pronunciation variants. *Marco van de Ven*
66. Speech reduction in spontaneous French and Spanish. *Francisco Torreira*
67. The relevance of early word recognition: Insights from the infant brain. *Caroline Junge*
68. Adjusting to different speakers: Extrinsic normalization in vowel perception. *Matthias J. Sjerps*

69. Structuring language. Contributions to the neurocognition of syntax. *Katrien R. Segaert*
70. Infants' appreciation of others' mental states in prelinguistic communication: A second person approach to mindreading. *Birgit Knudsen*
71. Gaze behavior in face-to-face interaction. *Federico Rossano*
72. Sign-spatiality in Kata Kolok: how a village sign language of Bali inscribes its signing space**.** *Conny de Vos*
73. Who is talking? Behavioural and neural evidence for norm-based coding in voice identity learning. *Attila Andics*
74. Lexical processing of foreign-accented speech: Rapid and flexible adaptation. *Marijt Witteman*
75. The use of deictic versus representational gestures in infancy. *Daniel Puccini*
76. Territories of knowledge in Japanese conversation. *Kaoru Hayano*
77. Family and neighbourhood relations in the mental lexicon: A cross-language perspective. *Kimberley Mulder*
78. Contributions of executive control to individual differences in word production. *Zeshu Shao*
79. Hearing speech and seeing speech: Perceptual adjustments in auditory-visual processing. *Patrick van der Zande*
80. High pitches and thick voices: The role of language in space-pitch associations. Sarah Dolscheid
81. Seeing what's next: Processing and anticipating language referring to objects. *Joost Rommers*
82. Mental representation and processing of reduced words in casual speech. *Iris Hanique*
83. The many ways listeners adapt to reductions in casual speech. *Katja Poellmann*
84. Contrasting opposite polarity in Germanic and Romance languages: Verum Focus and affirmative particles in native speakers and advanced L2 learners. *Giuseppina Turco*
85. Morphological processing in younger and older people: Evidence for flexible dual-route access. *Jana Reifegerste*
86. Semantic and syntactic constraints on the production of subject-verb agreement. *Alma Veenstra*
87. The acquisition of morphophonological alternations across languages. *Helen Buckler*
88. The evolutionary dynamics of motion event encoding. *Annemarie Verkerk*
89. Rediscovering a forgotten language. *Jiyoun Choi*

90. The road to native listening: Language-general perception, language-specific input. *Sho Tsuji*
91. Infants'understanding of communication as participants and observers. *Gudmundur Bjarki Thorgrímsson*
92. Information structure in Avatime. *Saskia van Putten*
93. Switch reference in Whitesands. *Jeremy Hammond*
94. Machine learning for gesture recognition from videos. *Binyam Gebrekidan Gebre*
95. Acquisition of spatial language by signing and speaking children: a comparison of Turkish sign language (TID) and Turkish. *Beyza Sümer*
96. An ear for pitch: on the effects of experience and aptitude in processing pitch in language and music. *Salomi Asaridou*
97. Incrementality and flexibility in sentence production. *Maartje van de Velde*
98. Social learning dynamics in chimpanzees: Reflections on (nonhuman) animal culture. *Edwin van Leeuwen*
99. The request system in Italian interaction. *Giovanni Rossi*
100. Timing turns in conversation: A temporal preparation account. *Lilla Magyari*
101. Assessing birth language memory in young adoptees. *Wencui Zhou*
102. A social and neurobiological approach to pointing in speech and gesture. *David Peeters*
103. Investigating the genetic basis of reading and language skills. *Alessandro Gialluisi*