Behavioral/Cognitive

Toward Robust Functional Neuroimaging Genetics of Cognition

Julia Uddén,[1,2,3,4] Annika Hultén,[1,2] Katarina Bendtz,[4] Zachary Mineroff,[5] Katerina S. Kucera,[1] Arianna Vino,[1] Evelina Fedorenko,[5,6,7] Peter Hagoort,[1,2] and Simon E. Fisher[1,2]

[1]Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands, 6525 XD, [2]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, the Netherlands, 6500 HE, [3]Department of Linguistics, [4]Department of Psychology, Stockholm University, Sweden, SE-106 91, [5]Brain and Cognitive Sciences Department, Massachusetts Institute of Technology, Cambridge, Massachusetts, MA 02139-4307, [6]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, MA 02139, and [7]Psychiatry Department, Massachusetts General Hospital, Charlestown, Massachusetts MA 02144

A commonly held assumption in cognitive neuroscience is that, because measures of human brain function are closer to underlying biology than distal indices of behavior/cognition, they hold more promise for uncovering genetic pathways. Supporting this view is an influential fMRI-based study of sentence reading/listening by Pinel et al. (2012), who reported that common DNA variants in specific candidate genes were associated with altered neural activation in language-related regions of healthy individuals that carried them. In particular, different single-nucleotide polymorphisms (SNPs) of *FOXP2* correlated with variation in task-based activation in left inferior frontal and precentral gyri, whereas a SNP at the *KIAA0319/TTRAP/THEM2* locus was associated with variable functional asymmetry of the superior temporal sulcus. Here, we directly test each claim using a closely matched neuroimaging genetics approach in independent cohorts comprising 427 participants, four times larger than the original study of 94 participants. Despite demonstrating power to detect associations with substantially smaller effect sizes than those of the original report, we do not replicate any of the reported associations. Moreover, formal Bayesian analyses reveal substantial to strong evidence in support of the null hypothesis (no effect). We highlight key aspects of the original investigation, common to functional neuroimaging genetics studies, which could have yielded elevated false-positive rates. Genetic accounts of individual differences in cognitive functional neuroimaging are likely to be as complex as behavioral/cognitive tests, involving many common genetic variants, each of tiny effect. Reliable identification of true biological signals requires large sample sizes, power calculations, and validation in independent cohorts with equivalent paradigms.

Key words: **fMRI; FOXP2; individual differences; KIAA0319/TTRAP/THEM2; language; neuroimaging genetics**

Significance Statement

A pervasive idea in neuroscience is that neuroimaging-based measures of brain function, being closer to underlying neurobiology, are more amenable for uncovering links to genetics. This is a core assumption of prominent studies that associate common DNA variants with altered activations in task-based fMRI, despite using samples (10–100 people) that lack power for detecting the tiny effect sizes typical of genetically complex traits. Here, we test central findings from one of the most influential prior studies. Using matching paradigms and substantially larger samples, coupled to power calculations and formal Bayesian statistics, our data strongly refute the original findings. We demonstrate that neuroimaging genetics with task-based fMRI should be subject to the same rigorous standards as studies of other complex traits.

Introduction

Advances in genomics are helping to identify genes contributing to key aspects of human cognition, including language (Deriziotis and Fisher, 2017). Insights come partly from rare high-impact mutations disrupting brain development (Fisher and Scharff, 2009). Nonetheless, even for highly heritable traits, the genetic architecture of interindividual variability in cognition and behavior mainly involves effects of many common variants

Received April 19, 2019; revised Aug. 21, 2019; accepted Sept. 4, 2019.

Author contributions: J.U., A.H., E.F., P.H., and S.E.F. designed research; J.U., A.H., Z.M., K.S.K., and A.V. performed research; J.U. and K.B. analyzed data; J.U. wrote the first draft of the paper; J.U., A.H., K.B., K.S.K., E.F., P.H., and S.E.F. edited the paper; J.U., K.B., E.F., and S.E.F. wrote the paper.

This work was supported by the Max Planck Society and the Donders Centre for Cognitive Neuroimaging. E.F. was supported by the National Institutes of Health (Grants R00-HD057522, R01-DC016607, and R01-DC016950) and by a grant from the Simons Foundation to the Simons Center for the Social Brain at MIT. E.F. would also like to acknowledge the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT and its

(polymorphisms) at different genomic loci. Typically, one common polymorphism has, at best, tiny effects on the relevant behavioral/cognitive phenotypes, so large cohorts of participants are needed to robustly detect genetic associations.

Neuroimaging enables noninvasive investigation of brain structure and function in living humans. A widely held view is that neuroimaging-derived measures, being closer to the underlying biology, are more amenable for tracing links to genetic variation than behavioral/cognitive assessments (Bigos et al., 2016; but see Flint and Munafò, 2007 for an opposing view). Twin studies using MRI-derived data (Thompson et al., 2001) suggest that a substantive proportion of interindividual variability in brain structure is explained by genetic variation in aggregate. However, systematic genome-wide studies indicate that genetic accounts of neuroanatomical variability are as complex as distal measures of behavior, such that effect sizes of individual polymorphisms remain small. For example, in screening of subcortical volumes in >30,000 individuals, the strongest association signal accounted for <0.52% of variance in putamen volume (Hibar et al., 2015).

Perhaps task-based functional MRI (fMRI), indexing neural activation related to particular cognitive processes, holds greater promise for bridging to genetics? Twin studies again support significant genetic contributions, although heritability estimates vary by task and/or brain region (Koten et al., 2009; Blokland et al., 2011). Whereas large genetic screens are feasible for structural MRI, scaling up task-based fMRI is difficult, with most studies limited to tens of individuals. Lacking power for genome-wide scans, task-based fMRI genetics efforts target candidate genes, yielding claims of positive associations with different polymorphisms for various paradigms (Grabitz et al., 2018). However, small cohorts in studies of complex traits have not only reduced power but also elevated false-positive rates (Button et al., 2013). Moreover, functional neuroimaging genetics involves multidimensional datasets, with multiple flexible parameters when processing/analyzing primary data, dramatically increasing the degrees of freedom (Simmons et al., 2011). This raises risks of p-hacking and HARKing (hypothesizing after results are known) (Grabitz et al., 2018). Few positive genetic associations with task-based fMRI have been formally replicated using matching designs; there are no direct replications in the language processing literature.

Here, we address this gap by focusing on the most highly cited neuroimaging genetics study of language function (Pinel et al., 2012) (179 Google Scholar citations at time of writing, corresponding to ~26 citations per year). Pinel et al. (2012) correlated single-nucleotide polymorphisms (SNPs) with task-based activation in language-related brain areas in 94 healthy participants, targeting two candidate loci from genetic investigations of language-related disorders: *FOXP2* and *KIAA0319/TTRAP/THEM2*. Rare *FOXP2* mutations cause a monogenic disorder, involving speech apraxia, expressive/receptive language impairments (Lai et al., 2001; Fisher and Scharff, 2009), and distributed alterations in brain structure and function (Watkins et al., 2002; Liégeois et al., 2003). The *KIAA0319/TTRAP/THEM2* locus was selected from genetic studies of dyslexia; clusters of common SNPs in this region have been associated with variation in reading skills (reviewed by Carrion-Castillo et al., 2013). On analyzing 39 SNPs from *FOXP2* and *KIAA0319/TTRAP/THEM2*, Pinel et al. (2012) found that three were associated with altered activation during covert sentence reading, each with a distinct pattern (Fig. 1). rs6980093 and rs7784315 (*FOXP2* SNPs) were correlated with variable activation in left inferior frontal and precentral gyri respectively. rs17243157 (*KIAA0319/TTRAP/THEM2*) was associated with variation in functional asymmetry of the superior temporal sulcus.

The present study attempted a hypothesis-driven replication of these findings. With 427 healthy participants, a sample four times that of Pinel et al. (2012), we believe this to be the largest fMRI genetics investigation of language-related activation to date. We demonstrate nonreplication of all three findings of the prior report.

Materials and Methods

Cohorts

The 427 participants in the present study came from two independently collected cohorts in which language task-based fMRI testing was coupled with DNA collection: referred to here as MOUS (Mother of Unification Studies) and EvLab. The MOUS cohort comprised 217 participants, 107 of whom performed a sentence reading (visual language) task and 110 performed a sentence listening (auditory language) task. The EvLab cohort comprised 210 participants, all of whom performed a visual language task. Table 1 summarizes the fMRI design and comparison to the Pinel et al. (2012) study, detailed further below.

Participants

MOUS cohort. A total of 242 native Dutch-speaking participants volunteered to participate in MOUS, an in-depth multimethod neuroimaging study of language processing, performed at the Donders Centre for Cognitive Neuroimaging (DCCN) in Nijmegen, the Netherlands. All participants completed an fMRI and a MEG session. Of those, 25 participants were excluded for a range of reasons, including: technical problems during data collection; poor data quality due to excessive blinking during initial MEG measurement, leading to exclusion from subsequent fMRI sessions; failure to complete the study; noncompliance with task instructions assessed via a behavioral threshold (see task description below). Approximately half of the remaining 217 participants read sentences and word lists, presented word-by-word (visual group: 107 participants; 51% male; mean age of 22.3 years, age range 18 to 33 years). The other half listened to auditory versions of the same materials (auditory group: 110 participants; 46% male; mean age of 22.3 years, age range 18 to 30 years). We will refer to the "cohorts" (to distinguish the MOUS and the EvLab sample) as well as the "subsamples" (to distinguish the visual MOUS sample, the auditory MOUS sample and the EvLab cohort). Key design features of the cohorts are noted in Table 1. All participants were right-handed as assessed by the Bever handedness questionnaire (building on Oldfield, 1971), had normal or corrected-to-normal vision, and reported no history of neurological, developmental or language deficits. Participants were instructed to not use medication, alcohol or drugs on the day of measurement. The study was approved by the local ethics committee (CMO, the local "Committee on Research Involving Human Subjects" in the Arnhem-Nijmegen region) and followed the guidelines of the Helsinki declaration.

EvLab cohort. The EvLab cohort was collected by Evelina Fedorenko's laboratory at the Massachusetts Institute of Technology (MIT) in Boston. This cohort consisted of 210 native English-speaking participants (36% male; mean age of 26.9 years, age range 20 to 59 years) who were students at MIT and members of the larger Boston community. They participated for payment in different neuroimaging experiments that all included a language localizer task (Fedorenko et al., 2010). As in the visual subset of the MOUS cohort, the EvLab participants read sentences presented word-by-word, as well as one or more control conditions, like lists of words and/or nonwords (Table 1). All participants were right-

support team (Steve Shannon and Atsushi Takahashi). We thank Matt Siegelman at EvLab and MOUS team members Karl Magnus Petersson, Jan-Mathijs Schoffelen, and Nietzsche Lam for contributions.

The authors declare no competing financial interests.

Correspondence should be addressed to Julia Uddén at Julia.Udden@mpi.nl or Simon E. Fisher at simon.fisher@mpi.nl.

https://doi.org/10.1523/JNEUROSCI.0888-19.2019

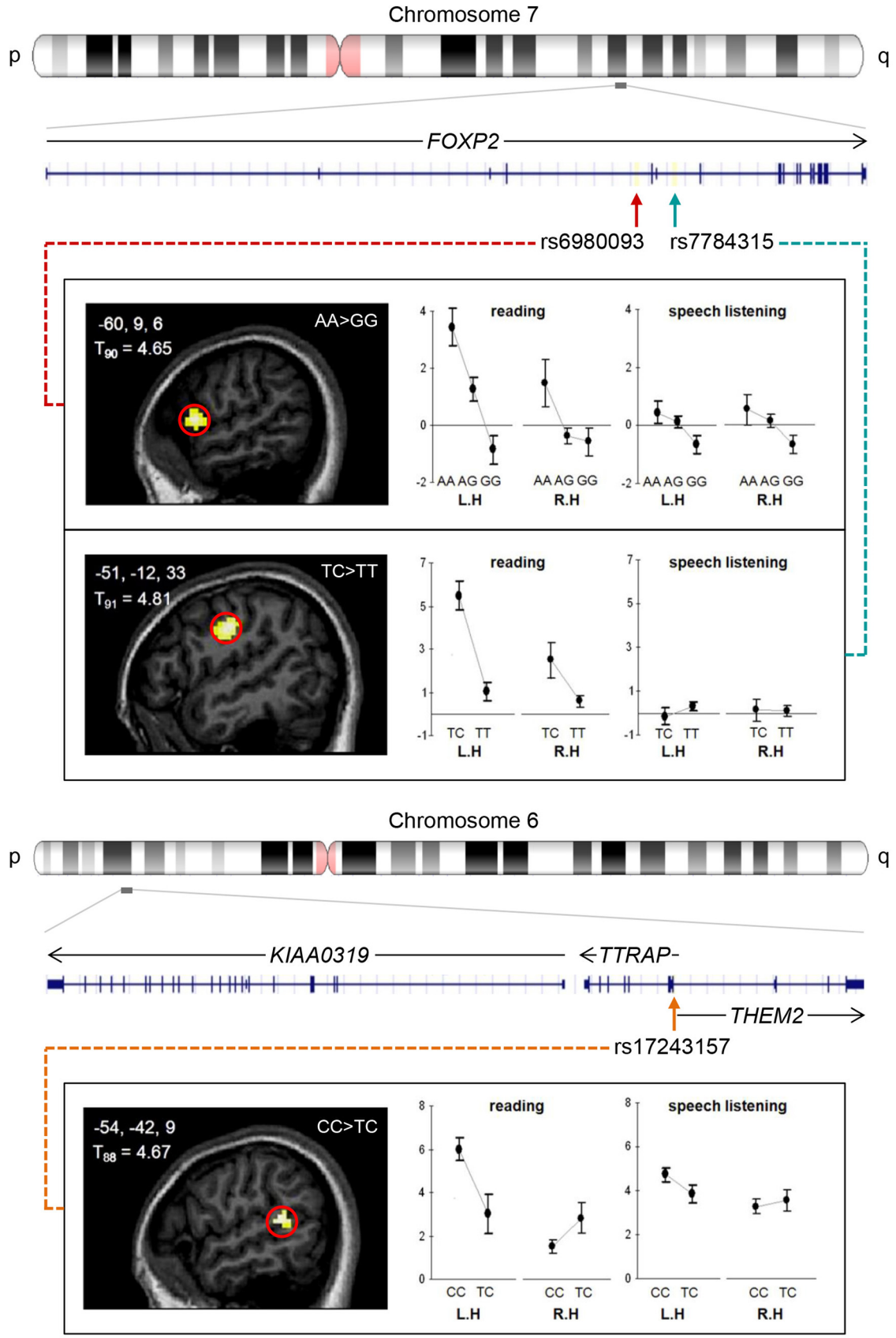

Figure 1. Hypotheses in the present study were directly based on prior findings of Pinel et al. (2012), displayed here (adapted from Fig. 2 of the original paper). *FOXP2* spans ~610 kilobases of chromosome 7q31.1. First, Pinel et al. (2012) associated rs6980093 and rs7784315 allelic status with reading-related activation of regions within the left inferior frontal gyrus and left precentral gyrus, respectively, as highlighted in the sagittal anatomical slices of the left panels. The right panels show the reported percentage changes in BOLD signal at cortical peaks of the left hemisphere (L.H) and the symmetrical position in the right hemisphere (R.H), plotted separately for the different allelic groups for the reading task. Following up on this association, Pinel et al. (2012) analyzed the corresponding speech-listening data, with results displayed to the very right. *KIAA0319*, *TTRAP*, and *THEM2* are neighboring genes spanning ~160 kilobases (*Figure legend continues.*)

Table 1. Summary of the key design features for the original study and the two cohorts investigated in the present report

Cohort (participants)	Sentences	Design	Compliance task(s)	Other conditions*
Visual/auditory from Pinel et al. (2012) ($N = 94$)	10 Visual, 10 auditory	Event related (10 trials per condition)	No task, readability of sentences assessed in postscan debriefing	Checkerboard view, simple button press, subtraction task
Visual ($N = 107$) and auditory ($N = 110$) from MOUS (present study)	60	Blocked, with five sentences per block (12 blocks per condition)	Comprehension question on 10% of the trials	Word lists
Visual ($N = 210$) from EvLab (present study)	36 – 48	Blocked, with three to five sentences per block (16 blocks per condition)	Memory probe task or simple button press task	Nonword lists, word lists

For all cohorts, the critical contrast examined was sentences versus a low-level control condition.

*Presented in the same fMRI experiment and modelled as separate events/blocks, but not included in the contrasts examined in the original/current study.

handed as assessed by the Oldfield (1971) handedness questionnaire or by self-report. All participants gave informed consent in accordance with the requirements of the MIT's Committee on the Use of Humans as Experimental Subjects (COUHES).

Language tasks

Both the MOUS and the EvLab cohort used a blocked design, in contrast to the Pinel et al. (2012) study, which relied on an event-related design (Table 1). In a blocked design, several stimuli from the same condition appear in a row (forming a block), and the response is estimated to the entire block. Given the additive nature of the BOLD signal, blocked designs have greater sensitivity (Birn et al., 2002) and are generally recommended over event-related designs except in cases where local predictability of experimental events could be problematic or where it is critical to examine responses to individual experimental events. Given that the primary goal of both the current study and of Pinel et al. (2012) was to reliably identify language-responsive cortical areas, so that properties of these areas could be related to genetic variation, a blocked design provides additional sensitivity.

Specifics of the paradigms have been reported in detail previously (Lam et al., 2016; Mahowald and Fedorenko, 2016; Schoffelen et al., 2017; Hultén et al., 2019; Uddén et al., 2019). Further specifics of the MOUS cohort are present in a peer reviewed data publication (Schoffelen et al., 2019). To check for compliance, participants in the MOUS cohort were presented with yes/no comprehension questions on 10% of the sentence trials. In the EvLab cohort, participants were asked to press a button at the end of each sentence, or to respond to a memory probe task (Fedorenko et al., 2010; Scott et al., 2017; it has been previously established that the nature of the task does not affect the resulting activation maps).

MRI acquisition

For both cohorts, structural and functional data were acquired with a SIEMENS Trio 3 T scanner (the MOUS cohort at the DCCN of the Donders Institute in Nijmegen and the EvLab cohort at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT). The functional scans used a TR of 2000 ms in both cases. Further specifics of the acquisition paradigms have been reported in detail previously (Mahowald and Fedorenko, 2016; Uddén et al., 2019).

Preprocessing and first-level analysis

Preprocessing and first-level analysis procedures were matched as closely as possible to those used in Pinel et al. (2012), to maximize the comparability of the results. We used the statistical parametric mapping software (SPM8; Wellcome Trust Centre for Neuroimaging, London; www.fil.ion.ucl.ac.uk/spm), including realignment to correct for individual subject head movement. Thus, one of the most critical sources of noise affecting the BOLD signal (i.e., subject motion) was controlled for in a similar manner across cohorts. A second important noise source: bodily physiology of the participant (e.g., heart rate), was not controlled for in any of the cohorts (Dubois and Adolphs, 2016). A correction for differences in slice acquisition time was performed for the MOUS cohort (ascending slice acquisition) but not for the EvLab cohort (interleaved slice acquisition). Structural images were spatially normalized to the structural image (T1) template provided by SPM8, using affine regularization. In this step, to match the original Pinel et al. (2012) study, the voxel size was resampled to $3.0 \times 3.0 \times 3.0$ mm^3. The transformation matrices generated by the normalization algorithm were applied to the corresponding functional EPI-BOLD volumes, after coregistration. All structural and functional images were smoothed with an isotropic 3D spatial Gaussian kernel (FWHM = 5 mm) matching Pinel et al. (2012). For the single-subject fixed effect analyses, we modeled the six realignment parameters from the movement correction. As in the original study, we used the canonical HRF and its temporal derivative. For the MOUS cohort, comprehension questions were modeled separately. As an example, in a block where a question appeared after sentence #1, we modeled two events with sentences: sentence 1 and sentences 2–5. As in Pinel et al. (2012), we computed individual asymmetry maps from the activation maps by subtracting the right from the left hemisphere, voxel by voxel, after normalizing the right hemisphere onto the left hemisphere.

Second-level region of interest (ROI) analysis

Because we sought to directly replicate the key findings reported by Pinel et al. (2012), the ROIs for this study were based on the locations of the peak associations reported for each of the significant SNPs in that study (Fig. 1). Specifically, we used a 10 mm spherical ROI around each of the three peak coordinates described in Pinel et al.: in the left inferior frontal gyrus (rs6980093), left precentral gyrus (rs7784315), and superior temporal gyrus/sulcus (rs17243157). Our analytical approach matched that used for each SNP in the original study. Specifically, for rs6980093 and rs7784315, we assessed left hemisphere activation for the sentence condition (relative to the fixation baseline, as in Pinel et al., 2012), extracted the average (across voxels) BOLD response in each ROI of each participant using MarsBaR (http://marsbar.sourceforge.net/) and analyzed the resulting averages in SPSS. For rs17243157, we analyzed asymmetry maps for the sentences > fixation contrast, averaged the difference values (left minus right, as described above; extracted using the spm_read_vols and Imcalc functions of SPM) across voxels in the ROI of each participant, and analyzed the resulting averages in SPSS. Participants with a deviation larger than four times the interquartile range from the mean within a subsample (and ROI), were considered outliers. Thus, if a participant was considered an outlier with respects to one (or several) ROIs, they were removed from all tests (this concerned two MOUS auditory subsample participants).

To test for the presence of reliable above-baseline activity in these three ROIs at the group level, regardless of the potential modulation by genotype group, we performed one-sample t tests (against 0) at the second level, for each subsample. If there were significant voxels at $p < 0.05$

←

(*Figure legend continued.*) of chromosome 6p22.3. Pinel et al. (2012) reported that rs17243157 allelic status was associated with altered functional asymmetry indices in the superior temporal gyrus/sulcus during the reading task, for the region highlighted in the sagittal anatomical slice of the left panel. The right panel shows reported percentage changes in peak BOLD signal of each hemisphere for the allelic groups, for the reading and speech-listening tasks. ROI locations of the present study are indicated with red outlines on each brain slice.

Table 2. Allele and genotype frequencies used in the power analysis

SNP	MAF			
rs6980093	G=0.468	GG=0.219	AG=0.498	AA=0.283
rs7784315	C=0.092	CC=0.008	CT=0.167	TT=0.824
rs17243157	T=0.118	TT=0.014	CT=0.208	CC=0.778

(using FWE correction) within the spherical ROI, we report those voxels as indicators of group-level activity within a ROI.

Power analysis

We used G*Power (http://www.gpower.hhu.de/en.html) to calculate the sample size we would need to have 80% power to detect the effect sizes (Cohen's f^2 for F tests; Cohen's d for t tests) reported by Pinel et al. (2012). t tests (see "Analysis structure" section for further information) were taken to be one-tailed, since we aimed to replicate a finding in the specific direction. In addition, again assuming 80% power, we estimated the minimal effect size that could be detected in the present investigation, both for the total combined sample and each separate cohort. Power depends not only on the total number of participants but also on the relative sizes of the different genotype groups, in this case, determined by the genotype frequencies for each SNP. We used the established allele frequencies (Table 2) from the European subset of the 1000 Genomes data recorded in the dbSNP database (https://www.ncbi.nlm.nih.gov/projects/SNP/) and assumed Hardy–Weinberg equilibrium.

Genotyping

The participants' saliva was collected with Oragene DNA collection OG500 kits (DNA Genotek) and DNA was extracted according to manufacturer's instructions. This investigation was a hypothesis-driven study in which we selected a priori those SNPs from Pinel et al. (2012) that best represented the major conclusions of that earlier report (rs6980093, rs7784315, and rs17243157, as shown in Fig. 2 of Pinel et al., 2012, and in Fig. 1 of the present manuscript). Because we were focusing on three specific SNPs, we did not perform genome-wide genotyping of hundreds of thousands of markers on DNA-chips for this study. Rather, we directly genotyped only the three SNPs of interest, using a specialized technique that has high sensitivity and specificity, known as KASP (Kompetitive Allele-Specific PCR; developed by LGC Genomics). Targeted genotyping of each SNP of interest was performed using customised KASP assays and a Bio-Rad CFX96 real-time PCR thermocycler. Each 10 μl reaction contained 0.14 μl of 72× KASP primer mix, 5 μl of 2× KASP master mix with standard ROX concentration, and 2 μl of genomic DNA (diluted at 20 ng/μl). Thermocycling was performed as follows: after initial 15 min denaturation at 94°C, 10 cycles were run with 20 s denaturation at 94°C followed by 45 s of annealing/elongation starting at 61°C and decreasing by 0.6°C per cycle, followed by 30 cycles with 10 s denaturation at 94°C followed by 45 s of annealing/elongation at 55°C. When necessary, five further cycles were performed using the same parameters as the final 30 cycles from the initial run. Genotypes were called using the CFX96 Manager software (Bio-Rad). For each SNP assay, every sample was genotyped in duplicate, and we also included negative controls and positive controls (DNA samples for which genotypes were already known) for each allelic combination, allowing us to verify that genotype calls were reproducible and accurate. In the few cases where samples yielded genotype calls that were unclear with KASP, we used direct Sanger sequencing to resolve ambiguities. Moreover, we performed additional validations by Sanger sequencing of randomly selected samples of each genotype. Through these efforts, we successfully determined genotypes for all three SNP markers of interest in all 427 participants.

Statistical tests

The effects of genotype groups on the BOLD response during sentence processing were analyzed within each cohort separately and in a combined analysis across cohorts, with F tests, as follows. For rs6980093, the analysis was run with an F test comparing all three allelic groups: AA homozygotes, AG heterozygotes, GG homozygotes. For rs7784315 and rs17243157, the frequency of the minor allele was only ~10% (consistent with population data, as shown in Table 2), such that the homozygous minor allele groups had sample sizes that would be much too small for meaningful analyses (see "Results" for further information). Thus, for each of those SNPs, we used an F test comparing two independent means, corresponding to two allelic groups: the major allele homozygotes and the heterozygotes. In these models, we included main effects and two-way interactions with genotype group and the following two potential confounding factors. Sex was included as a potential confounding factor for both cohorts. As the EvLab cohort contained slight variations in the experimental paradigm used across participants, we also included experiment version ("experiment" for short) as a possible confounding factor in addition to sex, for this cohort. Experiments with ≤5 participants were excluded from the EvLab cohort in this analysis to ensure robustness of the statistical tests. Due to the inclusion of potential confounding factors in the models, any results reported on the main effect of genotype, which is our focus, cannot be explained by confounding factors (regardless of their potential significance). In reporting our results, we provide information on the direction of effects observed in the earlier Pinel et al. (2012) study. Note, however, that directionality was not constrained in any test in this analysis.

We performed an additional set of analyses in JASP (https://jasp-stats.org/, JASP Team, 2017, version 0.8.4) to assess the Bayesian evidence for or against the hypothesis presented in the Pinel et al. (2012) report. The effects of genotype group on the BOLD response during sentence processing were analyzed either with a Bayesian F test or t test, as follows. For rs6980093, the analysis was run with a Bayesian F test with three genotype groups, corresponding to the three allelic groups (AA, AG, GG). The H_0 corresponds to no effect of genotype group and the H_1 to any effect of genotype group. We report the Bayes Factor (BF_{01}) comparing these two hypotheses. Following standard guidelines for interpreting Bayes Factors, $BF_{01} > 1$ is considered support for H_0 and $BF_{01} < 1$ is considered support for H_1, although $0.3 < BF_{01} < 3$ is regarded as weak evidence. For rs7784315 and rs17243157, each involving two allelic groups as explained above, we ran Bayesian t tests comparing two independent means. Based on the effects reported by Pinel et al. (2012) the H_1 was TC>TT for rs7784315, and CC>TC for rs17243157. In these cases, the H_0 corresponds to no genotypic effect, or an effect in the opposite direction. We used the default Cauchy prior distributions in JASP. Bayes factors are reported and labeled as inconclusive/no evidence, weak, substantial, or strong according to the guidelines set out by Jeffreys (1961) and JASP. Note that the Bayesian analyses, as currently implemented in JASP, did not allow for inclusion of potential confounding factors in these models. Bayes factors can only be estimated for complete models—that is, all main effects and interactions must be considered together as the H_1. Thus, we used the initial SPSS analysis to first exclude effects of possible confounding factors and then used a model with allelic group as the only factor for the subsequent Bayesian analysis.

Analysis structure

The analytic approach we have described thus had three parts: a power analysis, a classical statistical analysis, and a Bayesian analysis. All analyses were performed with comparable sets of tests that were matched as closely as possible to each other. In this paragraph, we comment on the correspondence of tests across these parts. Because the confounding factors (sex and experiment version; see "Statistical tests" section) were not reported in the original study, the confounding factors were not included in the power analysis. The power analysis was performed assuming an F test for rs6980093 and t tests for rs7784315 and rs17243157, as for rs6980093 there are three allele groups and for the other two SNPs there are two. The same division between F- and t tests across SNPs would also have been relevant for the classical statistical analysis, unless confounding factors had been included. However, because we sought to control for confounding factors in the classical analysis, F tests were in that case used for all SNPs. In the Bayesian statistical analysis, we did not include possible confounding factors, mainly because Bayesian model selection procedure compares the fit of total models and does not allow for single factors to be evaluated. In addition, because possible confounding factors had been included in the classical analysis, we were able to verify that they did not confound the results in this first step and then leave them out of the Bayesian analysis.

Table 3. Genotyping results

	$FOXP2_{rs6980093}$			$FOXP2_{rs7784315}$			$KIAA0319\ldots_{rs17243157}$		
Subsample	Allele	# (♂)	P_{HWE}	Allele	# (♂)	P_{HWE}	Allele	# (♂)	P_{HWE}
MOUS visual ($N = 107$)	AA	35 (19)	0.28	TT	82 (47)	0.60	CC	91 (46)	0.66
	AG	57 (28)		TC	24 (8)		TC	15 (9)	
	GG	15 (8)		CC	1 (0)*		TT	1 (0)*	
MOUS auditory ($N = 110$)	AA	35 (17)	0.27	TT	82 (38)	0.52	CC	95 (44)	0.44
	AG	59 (25)		TC	25 (10)		TC	15 (7)	
	GG	16 (9)		CC	3 (3)*		TT	0 (0)*	
EvLab ($N = 210$)	AA	82 (28)	0.13	TT	162 (60)	0.95	CC	155 (54)	0.93
	AG	90 (36)		TC	45 (16)		TC	51 (22)	
	GG	38 (12)		CC	3 (0)*		TT	4 (0)*	

Groups homozygous for the major allele are reported first at the top of each cell. Asterisks mark groups that were excluded due to insufficient size to support robust neuroimaging genetic studies. *p*-values for tests of Hardy–Weinberg equilibrium (HWE) are shown.

Results

The effect sizes of the findings reported by Pinel et al. (2012) were relatively large (reporting Cohen's f^2 for F tests; Cohen's d for t tests): Cohen's $f^2 = 0.29$ for rs6980093, Cohen's $d = 1.04$ for rs7784315, and Cohen's $d = 1.00$ for rs17243157. Our power calculations indicated that a cohort of 39 (rs6980093), 44 (rs7784315), or 40 (rs17243157) participants would give 80% power to detect effects of this magnitude, assuming an α level of 0.05. In our current study, we used cohorts that are substantially larger than this. Moreover, to account for the potential inflation of effect sizes in the Pinel et al. (2012) study (e.g., due to the winner's curse, Button et al., 2013), we also estimated the minimum effect size that we could detect in our cohorts, both for the separate cohorts and for the entire sample. For rs6980093, the minimum effect size was a Cohen's f^2 of 0.07 (separate cohorts) or 0.03 (entire sample). For rs7784315 and rs17243157, the minimum effect sizes (Cohen's d) for separate cohorts were 0.48 and 0.43, respectively, and for the entire sample, 0.32 and 0.29, respectively.

We used direct targeted genotyping to successfully determine the allelic status of rs6980093, rs7784315, and rs17243157 in all 427 participants. The minor allele frequencies were consistent with population data and there was no deviation from Hardy–Weinberg equilibrium for these SNPs in our cohorts or any of the subsamples (Table 3). For rs6980093, there were sufficient numbers of participants in each of the three allelic groups. For rs7784315 and rs17243157, the low minor allele frequency meant that homozygotes for this allele were rare (consistent with prior studies); between 0 and 4 participants in each subsample. Hence, and in line with the analyses performed by Pinel et al. (2012), the minor allele homozygotes were excluded from the imaging genetics analyses.

For all brain ROIs in all subsamples, the presence of significant voxels in the ROI established that there was activity at the group level for the sentence reading/listening task, as expected from prior studies. For all three cohorts, group level activation maps (Pinel, MOUS) or probabilistic activation overlap maps (Evlab), for the sentence versus low level baseline contrast, have been published previously (Pinel et al., 2007; Uddén et al., 2019, Mahowald and Fedorenko, 2016). Visual inspection of these maps showed reasonably high consistency across subsamples, in particular in the areas of interest. Peak voxels for the ROIs had the following locations in MNI space: MOUS Visual, [−57 17 10], left inferior frontal gyrus, $T = 7.5$, $p < 0.001$, [−57 −16 40], left precentral gyrus, $T = 6.2$, $p < 0.001$, [−57 −37 4], left superior temporal gyrus/sulcus, $T = 15.9$, $p < 0.001$; MOUS Auditory, [−57 −1 4], left inferior frontal gyrus, $T = 6.49$, $p < 0.001$, [−45 −4 40], left precentral gyrus, $T = 4.2$, $p = 0.002$, [51 −40 1], left superior temporal gyrus/sulcus, $T = 11.8$, $p < 0.001$; EvLab, [−54 14 10], left inferior frontal gyrus, $T = 12.5$, $p < 0.001$, [−48 −4 37], left precentral gyrus, $T = 13.4$, $p < 0.001$, [−54 −40 4], left superior temporal gyrus/sulcus, $T = 23.6$, $p < 0.001$. Note that small differences in the precise locations of peak group-level voxels are to be expected across cohorts given the well established interindividual anatomical and functional variability (Tomaiuolo et al., 1999; Fedorenko et al., 2010; Frost and Goebel, 2012). Pinel et al. (2012) reported significant effects of the *FOXP2* SNP rs6980093 on left inferior frontal gyrus activity for both auditory and visual modalities, with activity increasing with the number of A alleles (i.e., AA>AG>GG) (Fig. 1). We did not replicate this finding in any of the subsamples (Fig. 2, Table 4). For rs7784315, another *FOXP2* SNP in a different part of the gene, Pinel et al. (2012) described effects in the left precentral gyrus, with TC participants showing significantly higher activity than TT participants, but the association was only seen for the visual modality (Fig. 1). We did not replicate this finding in either of the two visual subsamples (Fig. 2, Table 4). As an additional analysis that must be considered exploratory, since no effect was observed in the Pinel et al. (2012) study, we tested rs7784315 in the MOUS auditory subsample. This auditory analysis generated a nominally significant uncorrected *P*-value (TC>TT group; $p = 0.02$), but the finding did not survive the appropriate Bonferroni correction for multiple comparisons. Finally, for rs17243157, a SNP in the *KIAA0319/TTRAP/THEM2* locus, Pinel et al. (2012) reported effects on asymmetry of activation in the superior temporal sulcus, with CC participants showing higher asymmetry than TC participants, for both auditory and visual modalities (Fig. 1). Again, we did not replicate these associations in any of the subsamples (Fig. 2, Table 4).

We excluded the possibility that the sex of the participants might confound our results. In the MOUS cohort, sex*genotype interactions as well as sex as a main effect were nonsignificant for all SNPs and for both visual ($p \geq 0.14$ for rs6980093, $p \geq 0.35$ for rs7784315, $p \geq 0.27$ for rs17243157) and auditory subsamples ($p \geq 0.54$ for rs6980093, $p \geq 0.30$ for rs7784315, $p \geq 0.29$ for rs17243157). In the EvLab cohort, for the analysis of rs6980093, sex was a marginally significant factor ($F_{(1,201)} = 3.96$, $p = 0.05$), and all other factors, including sex*genotype interactions, were nonsignificant ($p \geq 0.48$). For the EvLab analysis of rs7784315, no confounding factors were significant ($p \geq 0.18$), while for the analysis of rs17243157 "experiment" was significant ($F_{(1,197)} = 6.63$, $p = 0.01$) and all other factors were nonsignificant ($p \geq 0.09$).

In a joint analysis of the MOUS and the EvLab cohorts, combining all the available data from >400 people, none of the findings from Pinel et al. (2012) were replicated (Table 4). In these

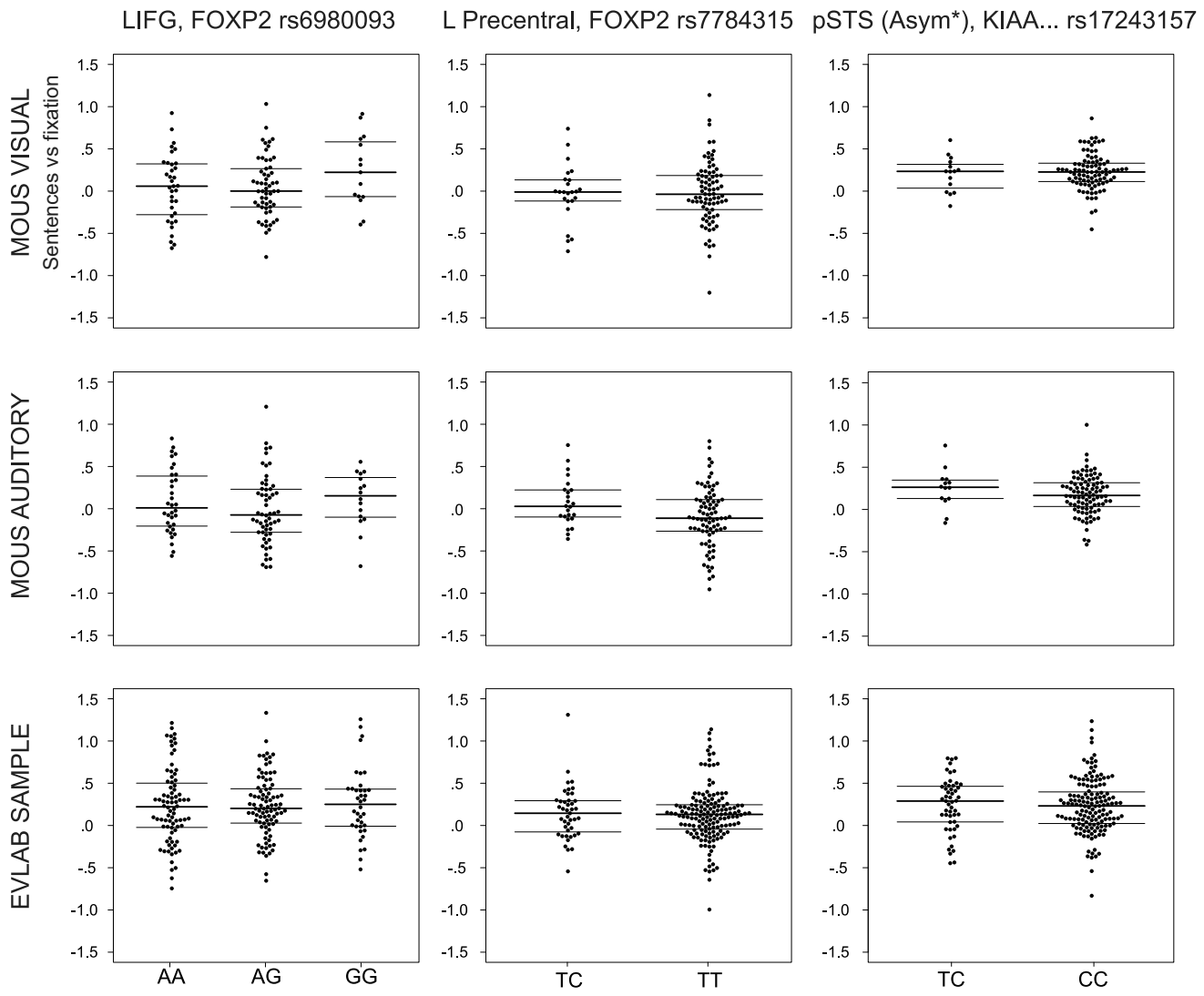

Figure 2. Single-subject contrast estimates by genotype group. The data points are horizontally jittered for visibility. The median as well as the first and third quartiles are shown as horizontal bars. For *FOXP2* rs6980093 in the original study, there was a negative relationship between the number of G alleles and the BOLD response (AA>AG>GG) to reading or listening to sentences in left inferior frontal gyrus. This finding was not replicated in either modality in the MOUS cohort, nor in the EvLab cohort. For *FOXP2* rs7784315 in the original study, the TC group had higher BOLD response in the left precentral gyrus when reading sentences, which is not replicated in any subsample in the current study. The corresponding results for the MOUS auditory subsample are reported for completeness, although we stress that there was no significant effect in the original study. For rs1724315 at the *KIAA0319/TTRAP/THEM2* locus, the original results showed greater functional asymmetry of the posterior STS for the CC compared with the TC group in the visual subsample, which is also not replicated in any subsample in the current study. The corresponding results for the MOUS auditory subsample are reported for completeness.

Table 4. Nonreplication of each of the main findings from Pinel et al. (2012) in multiple independent samples and in a combined analysis using *F* tests to assess genotype associations with ROI activations

$GENE_{SNP}$	Pinel et al. (2012) effects	Subsample	*p*-value	*F*-test value
$FOXP2_{rs6980093}$	AA > AG > GG	MOUS visual	0.16	$F_{(1,106)} = 1.85$
	Left inferior frontal gyrus activation	MOUS auditory	0.36	$F_{(1,107)} = 1.05$
	Visual and auditory	EvLab visual	0.66	$F_{(1,201)} = 0.41$
		Combined	0.20	$F_{(2,416)} = 1.60$
$FOXP2_{rs7784315}$	TC > TT	MOUS visual	0.85	$F_{(1,104)} = 0.37$
	Left precentral gyrus activation	EvLab visual	0.91	$F_{(1,198)} = 0.01$
	Visual only	Combined	0.97	$F_{(1,304)} = 0.00$
$KIAA0319\ldots_{rs17243157}$	CC > TC	MOUS visual	0.56	$F_{(1,105)} = 0.34$
	Superior temporal sulcus functional asymmetry	MOUS auditory	0.21	$F_{(1,107)} = 1.58$
	Visual and auditory	EvLab visual	0.48	$F_{(1,197)} = 0.50$
		Combined	0.30	$F_{(1,411)} = 1.07$

For rs6980093, we tested the three allelic groups, whereas for rs7784315 and rs17243157, we compared major allele homozygotes with heterozygotes, as the homozygous minor allele groups were too small for a meaningful comparison (see main text).

Table 5. Bayesian analysis provides evidence of nonreplication

$GENE_{SNP}$	Test hypothesis	Subsample	BF_{01}	Robustness
$FOXP2_{rs6980093}$	Any effect of genotype group, e.g. AA>AG>GG	MOUS visual	2.96	Weak
		MOUS auditory	4.45	Substantial
		EvLab visual	17.21	Strong
	AA>AG	MOUS visual	4.87	Substantial
		MOUS auditory	1.48	Weak
		EvLab visual	4.60	Substantial
	AG>GG	MOUS visual	8.54	Substantial
		MOUS auditory	6.61	Substantial
		EvLab visual	6.52	Substantial
	AA>GG	MOUS visual	7.52	Substantial
		MOUS auditory	3.84	Substantial
		EvLab visual	5.18	Substantial
$FOXP2_{rs7784315}$	TC>TT	MOUS visual	4.41	Substantial
		EvLab visual	7.14	Substantial
$KIAA0319\ldots_{rs17243157}$	CC>TC	MOUS visual	2.42	Weak
		MOUS auditory	6.74	Substantial
		EvLab visual	6.40	Substantial

H_1 hypotheses were based directly on the key findings of the Pinel et al. (2012) study, as detailed in main text (see also Figure 1). Bayes factors (reported as BF_{01}) are shown for the results of testing these hypotheses against H_0 (no replication of the original study) using Bayesian univariate ANOVA (rs6980093) or Bayesian independent-samples *t* test (rs7784315, rs17243157). $BF_{01} > 10$ indicates strong evidence and $10 > BF_{01} > 3$ indicates substantial evidence for a nonreplication (see Materials and Methods).

analyses, we included sex and "experiment" (see "Statistical tests" section for definitions) in the models, with MOUS auditory and MOUS visual paradigms considered as two different experiment versions. Note, however, that "experiment" in the joint analysis reflects a combination of factors (such as the exact set of participants, slight differences in the materials/procedure, any subtle variations in the scanning environment that are time-dependent, etc.), rather than an effect of sensory input modality per se. For combined analysis of rs6980093 the effect of "experiment" was significant ($F_{(3,416)} = 5.61$, $p = 10^{-3}$), as well as the effect of sex ($F_{(1,416)} = 5.01$, $p = 0.03$), but the interactions sex*genotype and "experiment"*genotype were all nonsignificant ($p \geq 0.60$). For combined analysis of rs7784315 the effect of "experiment" was significant ($F_{(2,304)} = 5.78$, $p = 3 \times 10^{-3}$), but the effect of sex as well as the interactions sex*genotype and "experiment"* genotype were all nonsignificant ($p \geq 0.58$). For combined analysis of rs17243157, the effect of "experiment" was significant ($F_{(3,411)} = 3.50$, $p = 0.02$), but the effect of sex as well as the interactions sex*genotype and "experiment"*genotype were all nonsignificant ($p \geq 0.45$). Thus, we excluded the possibility that sex and experiment version were confounding factors that would contribute to assessment of genotypic effects.

Finally, we used Bayesian analyses to formally evaluate the robustness of our findings (Table 5). Almost all of the nonreplication results reported above were robustly informative in the Bayesian sense, providing overall substantial to strong evidence for nonreplication (Table 5). The exceptions were in the MOUS visual subsample, with Bayes Factors of 2.96 (i.e., just below the threshold value of three) for rs6980093 and 2.42 for rs17243157. As noted above, rs7784315 showed no effect on left precentral gyrus activation in the auditory modality in the original Pinel et al. (2012) study; their positive (TC>TT) findings were restricted to the visual task. In an exploratory Bayesian analysis of the MOUS auditory subsample we found no evidence for a TC>TT effect of rs7784315 on activation of this region ($BF_{01} = 0.30$).

Discussion

Here, we attempted to replicate an influential fMRI-genetics investigation of language processing; a field with no direct replications in the literature. Pinel et al. (2012) reported association between rs6980093 and rs7784315 (*FOXP2* SNPs) and strength of the BOLD response during sentence reading/listening, in the left inferior frontal gyrus and precentral gyrus, respectively. They also reported association between rs17243157 (in the *KIAA0319/TTRAP/THEM2* locus) and asymmetry of the BOLD response during sentence reading/listening, in the posterior superior temporal sulcus. Despite including a substantially larger sample, and a more powerful blocked design (Birn et al., 2002), we did not replicate the Pinel et al. (2012) associations. Formal Bayesian analyses yielded substantial-to-strong evidence supporting the null hypotheses for each SNP. The lack of association was consistent across two independent cohorts (MOUS and EvLab), each itself larger than the original cohort. Power calculations indicate that combined investigations of MOUS and EvLab had 80% power to detect effect sizes of considerably lower magnitude than those observed in Pinel et al. (2012). Nevertheless, in combined analyses (>400 participants), association tests were all nonsignificant, indicating unambiguous lack of replication.

Our results contribute to a growing appreciation of the small magnitude of effects in neuroimaging genetics of common DNA variants, demonstrating that this limitation extends to task-based fMRI of higher cognition. It is worth considering factors that might have contributed to observations of positive signals in Pinel et al. (2012).

First, the earlier study began with exploratory screening of 39 SNPs distributed across two chromosomal regions (*FOXP2* on 7q31, *KIAA0319/TTRAP/THEM2* on 6p22.3), testing every marker for association with three different brain ROIs (based on fMRI-studies of different language-related disorders), each investigated using multiple approaches to phenotype definition, including hemisphere-specific activations and fMRI-based asymmetry indices. With multiple-testing adjustment for the effective number of independent SNPs, the significant p-values from this screen were $P_{corr}(M_{eff}) = 0.0286$ for rs6980093 (left activation, inferior frontal gyrus), $P_{corr}(M_{eff}) = 0.0182$ for rs7784315 (left activation, precentral gyrus), and $P_{corr}(M_{eff}) = 0.0234$ for rs17243157 (functional asymmetry, posterior superior temporal sulcus) (Fig. 1 in Pinel et al. (2012)). These findings cluster around $p = 0.05$; none would survive further multiple-testing adjustment to account for investigating several brain ROIs, and exploring multiple phenotypic definitions.

Second, to validate effects suggested by their exploratory screen, instead of testing top SNPs and brain ROIs in independently phenotyped/genotyped samples, Pinel et al. (2012) performed further SPM analyses of the same marker data, focusing on the cortical locations that had shown peak association with those SNPs. Since this involved reanalyzing the same input data used in the initial screen, testing hypotheses that had emerged from analyzing those data, it is a foregone conclusion that significant SNP effects were observed for the designated cortical locations (Kriegeskorte et al., 2009).

Another often neglected issue for designing and interpreting neuroimaging genetics studies concerns the question of whether/how SNPs impact gene function. While there is considerable interindividual variation in DNA sequence in humans, spread throughout the genome, the majority of common SNPs have no effects on expression, regulation or activity of genes and their encoded proteins. Since much of the interindividual variation observed at a gene locus is functionally neutral, most SNPs used in association screening have no mechanistic relevance. Experimental efforts are underway in genomics, using cellular models and tissue samples, to pinpoint functional DNA variants against the huge background of nonrelevant noise, and to understand

how such functional variants alter expression, regulation and activity of gene products (Gasperini et al., 2016; GTEx Consortium, 2017). For association studies in cohorts with low power, when multiple-testing becomes a limiting factor, evidence from functional experiments should ideally constrain SNP choice, before association testing. Pinel et al. (2012) only considered biological relevance as a *post hoc* discussion point, after observing which SNPs showed putative genotype-phenotype correlations.

The *FOXP2* locus is large (>610 kilobases) with many common SNPs in its introns, noncoding parts that are spliced out of mRNA transcripts before translation into protein (Hoogman et al., 2014). Unlike our mechanistic knowledge concerning rare disruptions (Sollis et al., 2017), we have limited understanding of which common SNPs in *FOXP2* have biological impact on its functions (Becker et al., 2018). Pinel et al. (2012) acknowledged the absence of experimental evidence that rs6980093 and/or rs7784315 affect *FOXP2* functions; the authors speculated based on their approximate locations within the gene locus, but such arguments could be made *post hoc* for many different SNPs. Even now, potential molecular mechanisms, along with an explanation of why the two SNPs should affect distinct brain ROIs, remain elusive. For rs17243157 in the *KIAA0319/TTRAP/THEM2* locus, Pinel et al. (2012) noted that this marker tends to cosegregate with another SNP, rs9461045, that has been correlated with altered *KIAA0319* expression (Dennis et al., 2009). However, rs9461045 was one of the SNPs from Pinel et al.'s association screen that failed to show significant association with functional asymmetry.

Highly penetrant mutations disrupting *FOXP2* lead to disordered speech/language development, associated with subtle but detectable alterations in brain structure and function (Lai et al., 2001; Watkins et al., 2002; Liégeois et al., 2003; Fisher and Scharff, 2009). However, the largest targeted studies of common SNPs spanning the gene, in hundreds of people from the general population, report no association with interindividual differences in either neuroanatomy (Hoogman et al., 2014) or measures of language performance (Mueller et al., 2016). Beyond language, *FOXP2* was one of 12 significant loci in a large-scale genome-wide scan of susceptibility to attention deficit/hyperactivity disorder (20,183 cases; 35,191 controls), but the associated SNPs differ from those studied here (Demontis et al., 2019).

Multiple investigations have taken Pinel et al. (2012) as the primary motivation for investigating the highlighted SNPs for associations with questionnaire-based data or behavioral/cognitive tests, but results have not been compelling. In a study of 882 healthy undergraduates who completed schizotypal personality questionnaires, Crespi et al. (2017) reported that rs7799109 (a SNP that cosegregates with rs7784315) was associated with self-report items related to inner speech ($p = 0.048$) and speech fluency ($p = 0.049$), and with strength of handedness, independent of direction ($p = 0.027$). A behavioral study of rs6980093 in ~200 people reported association with individual differences in accuracy and rate of speech-category learning (Chandrasekaran et al., 2015). However, the distribution of rs6980093 diploid genotypes in that study (proportions of people designated as AA, AG or GG) deviated markedly from the expected Hardy–Weinberg equilibrium (Chandrasekaran et al., 2015). This deviation contradicts all other studies involving rs6980093, as well as the information in public databases of worldwide genotypes, casting doubt on genotyping accuracy and undermining their association results.

In another targeted study, Mozzi et al. (2017) assessed rs6980093 in two Italian samples, reporting associations with semantic fluency in a population-based cohort (699 children; 3–11 years old), and with single-word-reading accuracy in 317 families ascertained for dyslexia (572 children; 6–18 years old). Putative allelic effects on these traits contradicted the additive findings of Pinel et al. (2012) (AA>AG>GG; Fig. 1) and were inconsistent between the Italian samples. In the population-based cohort, the G allele seemed dominant (AA<[AG=GG]), whereas the family cohort appeared to show heterozygous advantage (AG>[AA=GG]), although there is no mechanistic reason to favor this unusual mode of action for rs6980093 (Mozzi et al., 2017). Most recently, Zhang et al. (2018) tested 133 Chinese adults for behavioral and event-related potential responses (N1 and P2) to −50 and −200 cents pitch perturbations during vocal production, reporting that rs6980093 GG carriers performed differently from AA and AG carriers, but only for aspects of the −200 condition. Notably, until now, no study of the SNPs of interest attempted direct replication of Pinel et al. (2012); even when SNP choice was the same, different phenotypes were targeted. Moreover, none have reported fMRI associations.

Although unlikely, we acknowledge that lack of replication in the present study could partly reflect specific aspects of study materials, tasks, and/or cohorts that differed from Pinel et al. (2012). Most participants in both studies were young educated Caucasians, with similar age and sex distributions. One difference between the original design and that of our MOUS cohort, was use of a comprehension task in the latter for ensuring compliance. However, we also find strong nonreplication in EvLab, where memory probes or simple button-press tasks were used. Furthermore, prior work has established that language activations in high-level areas are robust to changes in materials, presentation modality, language, and task (Fedorenko et al., 2010; Scott et al., 2017). We note that if putative genetic effects are real, but so sensitive to subtle aspects of study design that they fail to replicate, then their relevance for understanding biological pathways that are fundamental for language function is minimal.

Future investigations in functional neuroimaging genetics should take into account the typically tiny effects of common polymorphisms when justifying sample size, include appropriate power calculations, address multiple testing robustly, and carefully consider functional relevance of targeted SNPs, given that most variants have no biological impact. Independent replications using directly matching study designs must be encouraged, so that the literature of this emerging field can be built on solid foundations.

References

Becker M, Devanna P, Fisher SE, Vernes SC (2018) Mapping of human FOXP2 enhancers reveals complex regulation. Front Mol Neurosci 11:47.

Bigos KL, Hariri AR, Weinberger DR (2016) Neuroimaging genetics: principles and practices. New York: OUP.

Birn RM, Cox RW, Bandettini PA (2002) Detection versus estimation in event-related fMRI: choosing the optimal stimulus timing. Neuroimage 15:252–264.

Blokland GA, McMahon KL, Thompson PM, Martin NG, de Zubicaray GI, Wright MJ (2011) Heritability of working memory brain activation. J Neurosci 31:10882–10890.

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14:365–376.

Carrion-Castillo A, Franke B, Fisher SE (2013) Molecular genetics of dyslexia: an overview. Dyslexia 19:214–240.

Chandrasekaran B, Yi HG, Blanco NJ, McGeary JE, Maddox WT (2015) Enhanced procedural learning of speech sound categories in a genetic variant of FOXP2. J Neurosci 35:7808–7812.

Crespi B, Read S, Hurd P (2017) Segregating polymorphisms of FOXP2 are

associated with measures of inner speech, speech fluency and strength of handedness in a healthy population. Brain Lang 173:33–40.
Demontis D, Walters RK, Martin J, Mattheisen M, Als TD, Agerbo E, Baldursson G, Belliveau R, Bybjerg-Grauholm J, Baekvad-Hansen M, Cerrato F, Chambert K, Churchhouse C, Dumont A, Eriksson N, Gandal M, Goldstein JI, Grasby KL, Grove J, Gudmundsson OO, et al. (2019) Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. Nat Genet 51:63–75.
Dennis MY, Paracchini S, Scerri TS, Prokunina-Olsson L, Knight JC, Wade-Martins R, Coggill P, Beck S, Green ED, Monaco AP (2009) A common variant associated with dyslexia reduces expression of the KIAA0319 gene. PLoS Genet 5:e1000436.
Deriziotis P, Fisher SE (2017) Speech and language: translating the genome. Trends Genet 33:642–656.
Dubois J, Adolphs R (2016) Building a science of individual differences from fMRI. Trends Cogn Sci 20:425–443.
Fedorenko E, Hsieh PJ, Nieto-Castañón A, Whitfield-Gabrieli S, Kanwisher N (2010) New method for fMRI investigations of language: defining ROIs functionally in individual subjects. J Neurophysiol 104:1177–1194.
Fisher SE, Scharff C (2009) FOXP2 as a molecular window into speech and language. Trends Genet 25:166–177.
Flint J, Munafò MR (2007) The endophenotype concept in psychiatric genetics. Psychol Med 37:163–180.
Frost MA, Goebel R (2012) Measuring structural–functional correspondence: spatial variability of specialised brain regions after macro-anatomical alignment. Neuroimage 59:1369–1381.
Gasperini M, Starita L, Shendure J (2016) The power of multiplexed functional analysis of genetic variants. Nat Protoc 11:1782–1787.
Grabitz CR, Button KS, Munafò MR, Newbury DF, Pernet CR, Thompson PA, Bishop DVM (2018) Logical and methodological issues affecting genetic studies of humans reported in top neuroscience journals. J Cogn Neurosci 30:25–41.
GTEx Consortium (2017) Genetic effects on gene expression across human tissues. Nature 550:204–213.
Hibar DP, Stein JL, Renteria ME, Arias-Vasquez A, Desrivières S, Jahanshad N, Toro R, Wittfeld K, Abramovic L, Andersson M, Aribisala BS, Armstrong NJ, Bernard M, Bohlken MM, Boks MP, Bralten J, Brown AA, Chakravarty MM, Chen Q, Ching CR, et al. (2015) Common genetic variants influence human subcortical brain structures. Nature 520:224–229.
Hoogman M, Guadalupe T, Zwiers MP, Klarenbeek P, Francks C, Fisher SE (2014) Assessing the effects of common variation in the FOXP2 gene on human brain structure. Front Hum Neurosci 8:473.
Hultén A, Schoffelen JM, Uddén J, Lam NHL, Hagoort P (2019) How the brain makes sense beyond the processing of single words: an MEG study. Neuroimage 186:586–594.
Jeffreys H (1961) Theory of probability, Ed 3. Oxford: Clarendon.
Koten JW Jr, Wood G, Hagoort P, Goebel R, Propping P, Willmes K, Boomsma DI (2009) Genetic contribution to variation in cognitive function: an FMRI study in twins. Science 323:1737–1740.
Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI (2009) Circular analysis in systems neuroscience: the dangers of double dipping. Nat Neurosci 12:535–540.
Lai CS, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP (2001) A forkhead-domain gene is mutated in a severe speech and language disorder. Nature 413:519–523.
Lam NHL, Schoffelen JM, Uddén J, Hultén A, Hagoort P (2016) Neural activity during sentence processing as reflected in theta, alpha, beta, and gamma oscillations. Neuroimage 142:43–54.
Liégeois F, Baldeweg T, Connelly A, Gadian DG, Mishkin M, Vargha-Khadem F (2003) Language fMRI abnormalities associated with FOXP2 gene mutation. Nat Neurosci 6:1230–1237.
Mahowald K, Fedorenko E (2016) Reliable individual-level neural markers of high-level language processing: a necessary precursor for relating neural variability to behavioral and genetic variability. Neuroimage 139:74–93.
Mozzi A, Riva V, Forni D, Sironi M, Marino C, Molteni M, Riva S, Guerini FR, Clerici M, Cagliani R, Mascheretti S (2017) A common genetic variant in FOXP2 is associated with language-based learning (dis)abilities: evidence from two Italian independent samples. Am J Med Genet B Neuropsychiatr Genet 174:578–586.
Mueller KL, Murray JC, Michaelson JJ, Christiansen MH, Reilly S, Tomblin JB (2016) Common genetic variants in FOXP2 are not associated with individual differences in language development. PLoS One 11:e0152576.
Oldfield RC (1971) The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9:97–113.
Pinel P, Thirion B, Meriaux S, Jobert A, Serres J, Le Bihan D, Poline JB, Dehaene S (2007) Fast reproducible identification and large-scale databasing of individual functional cognitive networks. BMC Neurosci 8:91.
Pinel P, Fauchereau F, Moreno A, Barbot A, Lathrop M, Zelenika D, Le Bihan D, Poline JB, 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.
Schoffelen JM, Hultén A, Lam N, Marquand AF, Uddén J, Hagoort P (2017) Frequency-specific directed interactions in the human brain network for language. Proc Natl Acad Sci U S A 114:8083–8088.
Schoffelen JM, Oostenveld R, Lam NHL, Uddén J, Hultén A, Hagoort P (2019) A 204-subject multimodal neuroimaging dataset to study language processing. Sci Data 6:17.
Scott TL, Gallée J, Fedorenko E (2017) A new fun and robust version of an fMRI localizer for the frontotemporal language system. Cogn Neurosci 8:167–176.
Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 22:1359–1366.
Sollis E, Deriziotis P, Saitsu H, Miyake N, Matsumoto N, Hoffer MJV, Ruivenkamp CAL, Alders M, Okamoto N, Bijlsma EK, Plomp AS, Fisher SE (2017) Equivalent missense variant in the FOXP2 and FOXP1 transcription factors causes distinct neurodevelopmental disorders. Hum Mutat 38:1542–1554.
Thompson PM, Cannon TD, Narr KL, van Erp T, Poutanen VP, Huttunen M, Lönnqvist J, Standertskjöld-Nordenstam CG, Kaprio J, Khaledy M, Dail R, Zoumalan CI, Toga AW (2001) Genetic influences on brain structure. Nat Neurosci 4:1253–1258.
Tomaiuolo F, MacDonald JD, Caramanos Z, Posner G, Chiavaras M, Evans AC, Petrides M (1999) Morphology, morphometry and probability mapping of the pars opercularis of the inferior frontal gyrus: an in vivo MRI analysis. Eur J Neurosci 11:3033–3046.
Uddén J, Hultén A, Schoffelen JM, Lam N, Harbusch K, van den Bosch A, Kempen G, Petersson KM, Hagoort P (2019) Supramodal sentence processing in the human brain: fMRI evidence for the influence of syntactic complexity in more than 200 participants. bioRxiv:576769.
Watkins KE, Vargha-Khadem F, Ashburner J, Passingham RE, Connelly A, Friston KJ, Frackowiak RS, Mishkin M, Gadian DG (2002) MRI analysis of an inherited speech and language disorder: structural brain abnormalities. Brain 125:465–478.
Zhang S, Zhao J, Guo Z, Jones JA, Liu P, Liu H (2018) The association between genetic variation in FOXP2 and sensorimotor control of speech production. Front Neurosci 12:666.