DISCOVERING GENES INVOLVED IN IMMUNITY, MITOCHONDRIA AND CILIA

INTEGRATIVE GENOMICS OF COMPLEX MOLECULAR SYSTEMS

ROBIN T.J.G. VAN DER LEE
The research presented in this thesis was performed at the Centre for Molecular and Biomolecular Informatics (CMBI), Radboud Institute for Molecular Life Sciences, Radboud university medical center, Nijmegen, The Netherlands.

The research was supported by the VIRGO consortium, funded by the Dutch government (project number FES0908), and by the Netherlands Genomics Initiative (project number 050-060-452).

Aerial image of a glacial river in Iceland. The river (data) consists of many interlinked streams with different colors and densities (heterogeneous data types). As the streams converge (integrate), they become more and more powerful. The research presented in this thesis is firmly based on the analysis and integration of patterns from large-scale molecular data, in order to better understand how life works.

The beauty of nature is inspiring, fragile, irreplaceable. Our species must do better to preserve it.

Cover image © Andre Ermolaev, used with permission (https://500px.com/andreabe)
Cover design Nurdianna Junaidi (http://nurdiannajunaidi.nl/)

Printed by Gildeprint
Design and layout Robin T.J.G. van der Lee
Copyright © 2016 Robin T.J.G. van der Lee
Additional licenses Chapters 2 and 6, © 2015 Van der Lee et al. Published under the Creative Commons Attribution 4.0 International License. (http://creativecommons.org/licenses/by/4.0/)
Chapter 3. © 2015 Springer-Verlag Berlin Heidelberg. Reproduced with permission from Springer.
Chapter 4. © 2016 American Academy of Allergy, Asthma & Immunology. Reproduced with permission from Elsevier.
Integrative Genomics of Complex Molecular Systems

Discovering genes involved in immunity, mitochondria and cilia

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen

in het openbaar te verdedigen op woensdag 23 november 2016
om 14.30 uur precies

door

Robin Theodorus Johannes Godefridus van der Lee

geboren op 27 augustus 1988
te Waalwijk
Promotor
Prof. dr. M.A. Huynen

Manuscriptcommissie
Prof. dr. G.J. Adema (voorzitter)
Dr. C.A. Albers
Prof. dr. V. van Noort (KU Leuven, België)
# Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
</table>
| Chapter 1 | General Introduction  
Outline of this thesis | 7  
32 |
| Chapter 2 | Integrative genomics-based discovery of novel regulators of the innate antiviral response | 35 |
|          | *PLOS Computational Biology* (2015), 11 (10): e1004553               |      |
| Chapter 3 | The RIG-I-like helicase receptor MDAS (*IFIH1*) is involved in the host defense against *Candida* infections | 75 |
|          | *European Journal of Clinical Microbiology & Infectious Diseases* (2015), 34 (5): 963-674 |      |
| Chapter 4 | Disease gene prioritization in recurrent herpes virus infections using systems immunology | 95 |
|          | *Journal of Allergy and Clinical Immunology* (2016), 138 (3): 895-898 |      |
| Chapter 5 | Genome-scale detection of positive selection in 9 primates predicts human-virus evolutionary conflicts | 101 |
|          | *Manuscript in preparation*                                         |      |
| Chapter 6 | Transcriptome analysis of complex I-deficient patients reveals distinct expression programs for subunits and assembly factors of the oxidative phosphorylation system | 129 |
|          | *BMC Genomics* (2015), 16 (1): 691                                    |      |
| Chapter 7 | CiliaCarta: an integrated and validated compendium of ciliary genes | 149 |
|          | *Submitted for publication*                                            |      |
| Chapter 8 | General Discussion                                                     | 159 |
| Chapter 9 | References  
Samenvatting in het Nederlands  
Summary  
Dankwoord / Acknowledgements  
Publications  
Curriculum Vitae | 168  
182  
185  
188  
192  
194 |
Chapter 1

General Introduction
General Introduction – Contents

Understanding life at the smallest level ................................................................. 9
Discovering the genes and proteins involved in a molecular system ....................... 9
A wealth of information for the modern computational biologist ......................... 11
Genomics data types and analyses ...................................................................... 11
  Genome sequences and comparative genome analysis ........................................ 12
  Gene expression data and co-expression analysis ............................................. 15
  Gene expression regulation and transcription factor binding site data ................. 17
  Protein interactions: within-species and between-species (host-pathogen) interactomes 21
  Functional data from systematic genetic screens .......................................... 23
  Genetics data and disease-associated variation ............................................. 24
Systematic descriptions of molecular systems through integrative genomics ............ 25
  Data integration: exploiting strengths and complementing weaknesses of individual data 25
  Statistics as a foundation for systematic discovery ....................................... 26
  Weighing evidence using probabilistic Bayesian integration .......................... 27
  Bioinformatics enrichment techniques for discovering properties of gene sets .... 28
Applications of integrative genomics for discovering components of molecular systems 28
  (i) Next-generation compendia of molecular system components .................. 29
  (ii) Global functional interaction networks of genes and proteins .................. 29
  (iii) Data-driven prioritization of candidate disease-causing variants and genes .... 30
  (iv) Other forms of genomics data combination ........................................... 30
Experimental validation to complement computational predictions ...................... 31
Outline of this thesis ......................................................................................... 32
Understanding life at the smallest level

How does life work? That question unites a broad range of scientific disciplines, each of which aims to shed light on a particular aspect of biology. To understand organisms at the microscopic level, we need knowledge about the molecular elements that make up their cells. Molecular biology primarily studies the DNA (our genes and genome), RNA and proteins, as these macromolecules are responsible for most of the functionality in the cell. Small molecules such as lipids, sugars and other metabolites are also important. One can ask many questions about our genes and proteins: What are their functions? How did they evolve? What do they look like, for instance in terms of their structure or sequence? How, when and where do they interact with other molecules? When and where are they present or absent? How is their activity regulated? Which cellular processes and systems do they take part in?

The interactions between tremendous numbers of different biomolecules give rise to complex molecular systems, which together give rise to cells, tissues, organs, organisms, and ultimately, to all of life. Thus, research into the small world of molecules improves our understanding of how life works. Fundamental research into molecular biology is a source of much innovation aimed at improving the quality of life. Applications include the development of biomaterials and the engineering of (micro)organisms and crops with desired characteristics that could for instance improve production of biofuels and food. A principal goal is to understand, diagnose, prevent, treat, and sometimes even cure disease, for example though the development of vaccines and drugs.

Discovering the genes and proteins involved in a molecular system

Genes and proteins that function together can be organized into biological pathways, networks and systems to provide context and facilitate studies into their combined roles. ‘Systems’ and ‘functions’ here can mean many things. For example, protein complexes formed by physical interactions between proteins (e.g. the proteasome or complex I of the oxidative phosphorylation system), cellular pathways that transduce energy, signals or other information through a network of cooperating proteins (e.g. pathogen detection pathways such as the antiviral RIG-I-like receptor pathway or the Toll-like receptor pathway), cellular substructures, such as organelles, with a specific protein composition (e.g. the mitochondrion or the cilium), or systems of disease genes in which mutations cause comparable phenotypes (e.g. susceptibility to infections, energy deficiencies, or ciliopathies). Molecular systems tend to be inherently complex in the sense that they consist of many parts that interact dynamically. The organization of these parts into larger systems results in emergent properties of systems that do not follow from the individual components alone.

Knowledge about the constituents of molecular systems and pathways is fundamental towards understanding their function in health and disease. Knowing which genes are involved in a particular system paves the way for further studies into, for example, (i) exactly how and where in the system these genes function, (ii) how they can be manipulated to alter the behavior of the system, (iii) why mutations in those genes cause diseases associated with that molecular system, and (iv) whether functional links may exist with other systems based on other known roles of the genes involved. For instance, knowing which genes are involved in the defense response against viral infections provides fundamental insights into how immunity works [1-3], offers clues as to
why mutations in such genes might cause recurrent infections or autoimmunity (e.g. through overstimulation of the immune system) \([4,5]\), could suggest points where viruses require interaction with host cells during infection (valuable for e.g. vaccine development) \([6,7]\), and may perhaps even suggest means for therapeutic intervention to boost (possibly reducing infection) or inhibit (possibly reducing autoimmunity) immune system functionality \([8,9]\).

This introduction provides an overview of recent advances in the systematic prediction and discovery of novel genes and proteins involved in molecular systems of interest (Figures 1 and 2). Much of this progress had been made possible by advances in (gen)omics methods, both experimental and computational, that systematically characterize the cell. To illustrate this, we will first discuss the modern availability of large-scale molecular data and describe in detail the basis and application of the most important data types: genome sequences, gene expression, transcription factor binding, protein interactions, functional screens, and genetic variation. Then we will introduce the field of ‘integrative genomics’, discuss the opportunities for bioinformatics data integration to maximally exploit the available data, and argue that computational approaches combined with experimental validation facilitate systematic discovery of the components of molecular systems. The final section will review successful cases where integration of large-scale molecular data has improved our description of molecular systems. Following this introductory chapter, the subsequent chapters of this thesis describe our studies that have exploited and advanced the data and techniques.
introduced here, to predict and identify novel genes involved in a variety of molecular and cellular systems important to human disease. Our studies focused on the innate antiviral and antifungal immune response, the virus-human interaction, the mitochondrial oxidative phosphorylation system, and the cilium (see Section ‘Outline of this thesis’ and Table 1).

A wealth of information for the modern computational biologist
Recent years have seen massive developments in the area of experimental techniques aimed at measuring the molecular states of cells with increasing detail and coverage [11-18]. Other fields such as particle- and astrophysics, economics, and climate science show parallel developments that lead to increasing availability of large-scale data [19-21]. In biology, among many other types of information we can now routinely obtain the sequences of whole genomes and exomes (the part of the genome that codes for proteins) of multiple species, populations and individuals; we can determine the repertoire of all expressed RNAs and proteins in many different types of cells and tissues; characterize genetic and physical interactions between genes and proteins under varying conditions; and obtain genome-wide information about the presence of transcription factors and how this affects gene expression. What unites such methods is their unprecedented breadth and depth. That is, the resulting ‘big data’ tend to (i) cover a large fraction or almost all of a specific type of information that could be potentially measured (i.e. these methods tend to be large-scale or even genome-scale), and (ii) they do not only measure which genes, proteins and interactions are present, but also determine amounts (i.e. they are often quantitative). The increasing availability of heterogeneous, large-scale data not only provides great opportunities for innovative computational analysis of the individual data sets, but also for approaches that combine various types of data, thereby unlocking novel ways for characterizing molecular systems.

Genomics data types and analyses
Throughout this thesis the total collection of large-scale data types describing the molecular state of cells is referred to as ‘genomics’ data. One could argue that, strictly, the term genomics would only apply to data that directly relates to genomic information such as genome sequence analysis; thus perhaps the term (gen)omics would be more accurate [11,22,23]. However, we here consider genomics data in the broadest sense to cover all aspects of large-scale or genome-scale molecular experiments and analyses, including any information describing the functions and interactions of genes and proteins.
The following sections provide an overview of the genomics data commonly used in modern large-scale analyses such as the studies presented in this thesis (Figure 2, Table 2), and for some data types discusses in more detail: (i) how much data is available and where it can be found, (ii) which (bioinformatics) methods are available for analyzing them and (iii) what specific insights can be derived from such analyses. Applications will be highlighted using examples from recent literature, primarily focusing on the human immune system. A large section will be devoted to the identification and applications of transcription factor binding sites. For descriptions of other molecular data types, see [11,12,14,16,24-26].

**Genome sequences and comparative genome analysis**

Genome sequencing involves the determination of the complete DNA sequence of an organism. Whole-genome and whole-exome sequencing has provided us with genome sequences of a large number of unique species [30,36-38]. For instance, the genomes of 43 mammals, among which are nine simian ('higher') primates, are publicly available from the Ensembl database (http://www.ensembl.org/info/about/species.html, release 83, December 2015) [27]. Given the rich availability of genome sequences, we can now compare these data in several ways to gain insight into the functions of genes and other genetic elements. The evolutionary analyses of genome sequences exploited in this thesis include applications to: (i) identify orthologous ('equivalent') genes in different species, (ii) predict genes functioning in the same biological system based on genomic presence-absence patterns, (iii) discover genes and codons showing signatures of adaptive evolution driven by positive selection, and (iv) identify transcription factor binding elements based on sequence conservation.

**Detecting orthologous genes for large-scale comparative analysis**

Orthologs are genes in different species that have evolved from a common ancestral gene following speciation [39]. In practice this concept is useful because orthologous genes tend to have equivalent functions in different species [40]. Therefore, orthology relationships are useful for annotating functions of unknown genes in newly sequenced genomes or otherwise poorly characterized sequences by transferring information from their orthologs, thereby implicitly assuming similar function. Similarly, practical reasons often lead researchers to study gene function...
in a model organism rather than in the actual species-of-interest, requiring accurate identification of the orthologous gene(s) in the model species [41]. Accurate orthology inference is also fundamental for more advanced evolutionary analyses comparing the ‘same’ gene across different species. For example, as discussed in more detail below, phylogenetic profiling works through annotating the presence/absence of orthologs in genomes, while the detection of positively selected genes depends on analyzing mutations across orthologous sequences.

Orthology relationships between genes are most accurately inferred from phylogenetic trees of homologous sequences [39]. These trees also provide clues about the evolutionary history of a gene family through events such as gene duplication and speciation. Phylogenetic analysis is a multi-step process requiring the identification of homologous protein sequences from different species (e.g. using BLAST-like methods), multiple sequence alignment, construction of the phylogenetic tree, and inference of evolutionary events. Resources such as PhylomeDB [42] and Ensembl Compara [43] have developed computational pipelines to automate these methods and apply them on a larger scale. Automatically obtained phylogenies are generally useful for bulk analysis, but individual cases can be unreliable due to a variety of challenges associated with large-scale sequence analysis [44]. In fact, this is true for most large-scale approaches, also outside the sequence domain. For example, the reliability of the protein sequences deposited into databases depends on accurate gene prediction, analysis of their intron/exon structure and potential transcripts, and translation of the genes into the likely RNA and protein sequences [45]. Other challenges include uncertainties in multiple sequence alignment [46] and phylogenetic tree reconstruction [44]. Furthermore, the
evolutionary history of large gene families with many duplications remains inherently more
difficult to dissect than, for example, that of clear one-to-one orthologs [39]. As one would expect,
the number of one-to-one orthologous relationships also decreases with increasing evolutionary
distances between genomes, which complicates comparisons between more divergent species
[40]. The large-scale analysis of positively selected genes in primates presented in Chapter
5 encountered several of the challenges associated with large-scale sequence analysis and
demonstrates that careful quality assessment and curation at each step of an automated sequence
analysis pipeline remain critical for obtaining high-confidence results.

Phylogenetic profiling of gene presence/absence to predict co-functioning genes
The gene content of genomes can be used to determine co-evolutionary presence-absence
relationships between orthologous groups across species, referred to as phylogenetic profiling
[47-49]. As genes functioning in the same biological system tend be gained and lost together
during evolution, such gene co-occurrence patterns can be correlated to each other to determine
systems of co-functioning genes (suggested by similar presence-absence patterns)[49], or mutually
exclusive pairs of genes with perhaps a similar or redundant function (one or the other gene
is present in a species)[50]. Phylogenetic profiles can also be correlated to a feature, such as a
phenotype, that is irregularly present across species to predict genes associated with that feature.
For example, the presence-absence pattern of the cilium organelle, a cellular substructure essential
for signaling and development, is highly predictive for ciliary genes: ciliated species tend to encode
ciliary genes, while non-ciliated species do not. This feature and several others were exploited in the
systematic construction of a compendium of ciliary genes (CiliaCarta, Chapter 7). Genes involved
in the specialized transition zone (TZ) of the cilium show an even more specific phylogenetic sub-
distribution that contributed to the identification of TMEM107 as a gene involved in the TZ [51].
Phylogenetic profiling has also successfully predicted components of small RNA pathways [52]
and mitochondria [53]. Resources are now emerging that allow systematic identification of genes
showing correlated genomic co-occurrence patterns with any feature or gene set of interest, such
as pathways, systems and phenotypes [54,55].

Identification of adaptively evolving genes in humans and primates
Genome sequences are also informative of adaptation and natural selection [56] within and
between species. Identification of the genes and processes that evolve rapidly can improve our
understanding of the evolutionary history of species in terms the selection pressure they faced
and what molecular adaptations underlie their specific phenotypes. For example, olfactory
receptor genes comprise one of the largest gene families in vertebrates with ~400 predicted
genes in human. However, olfactory genes show extreme diversity between species, which may
account for differences in smell perception [57]. Much evidence for molecular adaptation resulting
from positive selection comes from evolutionary analysis of protein-coding DNA sequences
from multiple species [58,59]. These methods detect genes and codons that show an excess of
nonsynonymous substitutions (mutations in the DNA that cause changes to the protein; non-
silent) compared to synonymous substitutions (mutations not changing the protein; silent) [60].
Chapter 5 provides a detailed explanation of $d_n/d_s$-based maximum likelihood methods, which are
the most popular and have been shown to achieve reliable results [61]. In Chapter 5, we performed
a large-scale comparative evolutionary analysis of nine whole-genome sequenced primates
(human, chimpanzee, gorilla, orangutan, gibbon, macaque, baboon, vervet and marmoset) to detect positively selected genes and investigated these genes in relation to antiviral immunity. The next part will introduce the phenomenon of rapid evolution of our immune system.

Selection and genetic adaptation in the human genome is a topic of intense study and speculation. Various molecular systems and functions have been shown to evolve rapidly [59,62-68]. Sensory perception, reproduction, apoptosis, morphology and possibly brain and speech development have been reported by some studies but not by others, depending on which genomes were compared and which detection methods and cutoffs were applied [59,66-68]. It has become clear however that genes associated with the host defense and immune response contain the most confident signatures of positive selection [62-65]. The rapid change of immune systems is driven by the continuously changing landscape and divergence of pathogens, which interfere with the immune response of their hosts [58,69,70]. Evolutionary analyses of primate sequences have previously revealed genetic conflicts between numerous primate proteins involved in the virus-host interaction and their interacting viral proteins [69]. These include TRIM5α [71], PKR [72], MxA [73], and mitochondrial antiviral signaling protein (MAVS), a signaling hub in the RIG-I-like receptor (RLR) pathway, which recognizes infections of a wide range of viruses from the presence of their RNA in the cytosol. In Chapter 2, we found that genes that are likely to have evolved under positive selection based on exome analysis of seven primates [63], are predictive for antiviral RLR pathway components. Going more in depth, our large-scale evolutionary analysis in primates (Chapter 5) identified 331 genes with signatures of positively selection (3% of human protein-coding genes). These 331 genes are strongly enriched for immune-related functions and formed the basis for an exploration of novel virus-human genetic conflicts through integration with for example virus-host protein interaction data and protein structures.

Conserved regulatory elements and transcription factor binding sites
Going beyond coding sequences, genome alignment of multiple species holds great power for detecting evolutionary conserved, and therefore probably functional, regulatory genetic elements such as transcription factor (TF) binding sites [74]. The availability of genomes for an increasing numbers of species has improved the reliability of conservation-based prediction of TF binding sites [30]. The section on ‘Gene expression regulation’ describes in more detail the identification of TF binding sites through sequence conservation and experimental methods, as well as the application of TF binding sites for discovering regulators of a molecular system.

Gene expression data and co-expression analysis
Large-scale profiling of gene expression through measurements of mRNA levels (transcriptomics) has long been possible using microarrays [75] and more recently using RNA sequencing [76]. Huge amounts of gene expression data are available from public databases such as the NCBI Gene Expression Omnibus, which currently contains over 1.6 million individual samples, of which almost 1 million are human, from over 62 thousand studies [28](http://www.ncbi.nlm.nih.gov/geo/summary/?type=history, December 2015).
Differential expression analysis

Gene expression data can be used to study numerous biological questions. For example, RNA sequencing not only provides expression levels but also offers insights into alternative splicing by determining which isoforms are present \[^78\] and can be used for expression profiling of small RNAs such as miRNAs in studies of post-transcriptional gene regulation \[^79\]. Clustering and comparison of expression levels offers insights into the differences between for instance health and disease (e.g. control and patient cells), developmental or differentiation states, the consequences of perturbations, environments and pathogens, etc. \[^80\]-\[^84\]. Furthermore, analysis of differentially expressed genes may point to common up- or down-regulated systems or pathways underlying the compared states. For example, genes that display substantial changes in their expression upon viral infection \[^85\] are highly predictive for components of the antiviral RLR pathway (part of the innate immune response, Chapter 2), while genes that are up-regulated during ciliogenesis induced by artificial removal of cilia from cells are highly enriched for known and novel ciliary genes \[^86\] (Chapter 7). Differential expression analysis of macrophages stimulated with the fungal pathogen *Candida albicans*, compared to unstimulated cells, revealed significant up-regulation of components of the antiviral RLR pathway, including the MDA5 receptor (Chapter 3).

Co-expression calculations

Gene pairs with correlated expression patterns (i.e. co-expressed genes) tend to be functionally related. This notion forms the basis of a large number of co-expression approaches that aim to identify genes that part are of the same molecular system. The signal for functional relationships in co-expression calculations has been shown to improve by including information from multiple species (conserved co-expression) \[^87\],\[^88\], by combining large amounts of gene expression data sets \[^89\], and by pre-selection or weighing of data sets that are the most predictive for a molecular system of interest \[^90\]-\[^92\]. The STRING database uses co-expression as one of the features for predicting functional interactions between genes alongside e.g. protein interaction data and genomic context (see Section ‘Applications of integrative genomics’\[^32\],\[^93\]). The recent two-step WeGET method (http://weget.cmbi.umcn.nl/) works by first systematically weighing ~1,000 expression data sets from mouse and human (~30,000 individual measurements) for their predictive potential by asking whether a group of query genes show high co-expression with each other \[^91\]. In step two, it integrates the data sets using the weighting scheme derived in step one to calculate for each gene a final conserved co-expression score with the query system \[^90\]. This approach has been effective for discovering genes involved in for example mitochondria \[^53\],\[^90\] and heme biosynthesis \[^94\], antiviral innate immunity (Chapter 2), the cilium (Chapter 7) and the ciliary transition zone \[^51\]. The study described in Chapter 6 analyzed RNA sequencing data of healthy and OXPHOS-deficient patient cell lines treated with various compounds that influence mitochondrial metabolism. Gene expression clustering of the various conditions, combined with (unweighted) co-expression calculations and TF binding site analyses indicated a different behavior of OXPHOS system subunits compared to assembly factors.

Investigating and defining cell types

Gene expression profiles are also valuable for characterizing cells and (re-)defining populations of cells with similar characteristics referred to as cell types \[^95\]. Definitions of cells types are often based on the expression of a limited number of (cell-surface) marker proteins, especially in the case
of the large variety of different immune cell subsets [96]. Compared to marker proteins, expression profiles either at the RNA or preferably the protein level provide a much more comprehensive view of what constitutes a particular type of cell or how cell types differ, leading to better understanding of what they can and cannot do. For example, the Immunological Genome Project determined transcriptomes of over 250 mouse immune cell types and stages of differentiation to investigate the hematopoietic process, how lineages of different cell types are related, and what distinguishes them [84]. On the single-cell level rather than the level of cell types or heterogeneous mixtures of cells, recent years have seen the development of RNA sequencing up to the point that high-resolution transcriptome analysis of individual cells is now possible [97].

**Gene expression regulation and transcription factor binding site data**

Transcription factors (TFs) are key controllers of gene expression. The human genome encodes an estimated ~1,500 TFs with specific DNA binding preferences (~6-7% of human genes), belonging to over 100 sequence families [98-100]. Precise spatio-temporal regulation of not only the expression and activity of the TFs themselves, but also e.g. the co-availability of interaction partners required for active transcription complexes are central to determining the expression landscape of a cell. In turn, expression patterns are what distinguishes cell types and tissues [83]. It has been proposed that a major part of phenotypic differences between individuals and species is the result of genomic variation influencing gene expression regulation, such as through divergence of transcription factor binding sites, rather than variation influencing gene content and coding regions [101]. Indeed, there are indications of extensive turnover and repurposing of TF binding sites between animal genomes [102,103].

**Transcription factor motifs and regulation of binding**

Transcription factors recognize the DNA at specific sequence stretches, referred to as binding motifs. Often, variations of the preferential motifs still allow binding to occur [104,105]. TFs from the same protein family or even from different families can have very similar binding specificities. For example, all members of the interferon regulatory factor (IRF1 to IRF9) family share a homologous DNA-binding domain, resulting in similar motif preferences for binding to the interferon response element (IRF-E; consensus sequence AANNGAAA) [106]. The redundancy in DNA binding preference between TFs does not mean that all TFs bind to all their potential binding motifs at all times. Specificity in gene regulation is achieved by (i) the small differences in DNA binding preferences that do still exist, but mostly by (ii) controlling the availability and activity of the TF, as dictated by expression levels, PTMs, and other mechanisms, (iii) cooperative binding with other factors, which may be present in one cell type or condition, but not in another, and (iv) the chromatin state of the genome [107,108]. Cooperative binding can even result in the same TF having apparent entirely distinct binding motifs [109].

TF DNA-binding motifs have been discovered and refined through various methods, including (i) high-throughput DNA binding essays, such as protein-binding microarrays [105,110], and (ii) bioinformatic *ab initio* motif discovery, for example in sequences found under ChIP peaks of TF binding [111,112]. JASPAR [113] and TRANSFAC [114] are two databases that collect curated TF regulatory motifs.
Transcription factor binding sites: experimental identification

Both experimental and computational approaches are valuable for identifying transcription factor binding sites [115]. Experimental data on genome-wide patterns of gene expression regulation have been obtained mainly using ChIP-seq and related next-generation sequencing methods that measure protein-DNA and chromatin interactions. DNase I hypersensitive site sequencing (DNase-seq) is a more indirect method for inferring potential TF binding [115]. It identifies open, accessible chromatin regions that are sensitive to cleavage by DNase I. These accessible regions often represent regulatory regions occupied by DNA-binding proteins. The NCBI Gene Expression Omnibus currently contains data from ChIP-seq, DNase-seq and related techniques from ~4,500 studies and ~82,000 individual samples (“Genome binding/occupancy profiling by high throughput sequencing”, http://www.ncbi.nlm.nih.gov/geo/summary/?type=history, November 2015). The ENCODE project has contributed a massive amount of TF binding and other functional genomics data, mapping the genomic locations of 119 different DNA-binding proteins in a maximum of 72 cell types when it was released in 2012 [29]. However, the vast majority of TFs have been assayed in one or a few conditions and cell types; primarily the three ‘tier 1’ cell lines GM12878 (B-lymphocyte, lymphoblastoid cell), H1-hESC (human embryonic stem cells), or K562 (an immortalized leukemia cell line). ENCODE currently stands at over one-thousand human ChIP-seq transcription factor data sets and over 350 DNase-seq data sets (https://www.encodeproject.org/search/?type=experiment, November 2015). These data can be browsed readily in the context of the genome using for example the UCSC Genome Browser [116]. Efforts are ongoing to measure the elements involved in gene expression regulation in increasingly more cell types and tissues [117].

Transcription factor binding sites: evolutionary conservation-based prediction

Comparative genomics approaches depend on evolutionary conservation for identifying potential TF binding sites (TFBS), thereby providing independent evidence that is complementary to ChIP-seq, other biochemical, and genetic approaches. Methods to predict TFBS first depend on the availability and reliability of the TF DNA-binding motifs. The next step in the computational identification of TFBS involves the scanning of genome sequences for the presence of motifs. However, due to the short length of the motifs, genome sequences by chance alone contain frequent motif matches, many of which are likely non-functional. Evolutionary conservation has been shown to greatly increase the power to discover functional TFBS [30]. Motifs that show a high level of sequence conservation across species, and preferably across species that have diverged over long evolutionary distances, may have evolved under purifying selection and therefore probably represent functional TFBS [74,118,119]. Despite accumulating evidence that functional TFBS indeed tend to be conserved, the concept does not always hold for at least two reasons [119]. (i) Conserved motifs need not always be functional, as for example fixation of neutral variants can also lead to apparently conserved motif instances without implying function, just like in coding sequences. (ii) Vice versa, motifs that appear non-conserved could indicate recent and ongoing innovations that may or may not indicate new functional binding sites. Challenges surrounding the correlation between conservation and function of TF binding sites, and other hurdles in
comparative genomics-based TFBS detection, such as the quality of multispecies whole-genome alignments, have been extensively discussed elsewhere [74,119,120].

The increasing availability of whole-genome sequences has enhanced the sensitivity and specificity of TFBS identification. For example, analysis of 29 mammalian genomes substantially improved the detection of individual functional motif instances compared to earlier studies using four, then-available species (human, mouse, rat, dog)[121]. The 29 mammals project generated 20 novel low-coverage genomes that together with nine previously completed genomes covered evolutionary divergence across the whole clade of placental mammals, such as primates, rodents and many farm animals [30]. In that study, conserved TF regulatory motifs were identified by calculating the smallest phylogenetic subtree that contains a motif match in human, as well as aligned motifs in other species, allowing for flexibility in the precise form and location of the motifs [122]. The level of conservation required for a TFBS to be included in the predictions depends on a confidence cutoff. TFBS identified at an estimated false discovery rate (FDR) of 60% showed reasonable agreement with binding sites obtained from ChIP-seq data, having a similar rate of biological relevance and overall reliability compared to TFBS from ChIP-seq data [30]. This thesis has made extensive use of the 29 mammals-based conserved TF binding sites (see Section ‘Predicting regulators of pathways and systems, below).

**Experimental and evolutionary approaches for TF binding site identification are complementary**

It is clear from the preceding discussion that TF binding sites obtained from biochemical experiments (e.g. the ENCODE project ChIP-seq data) and comparative sequence analyses (e.g. the 29 mammals evolutionary conservation data) are expected to differ. While both these TFBS data types have been explored to validate the quality of the other data type, they have intrinsic advantages and limitations that determine their suitability for answering different biological questions [74]. The main difference lies in the condition-specificity of experimental TFBS versus the total genome-wide potential of evolutionary conserved TFBS: individual ChIP-seq experiments measure TF binding in a specific cell type, under a specific condition, while conserved TFBS are usually detected solely on the basis of genome sequence. Despite the increasing amount of experimental TFBS information, only a subset of TFs have been measured for DNA binding, and those studies were inevitably limited to a subset of conditions. Even though aggregating all experimental TFBS data into a single data set should eventually deliver the total genome-wide potential of TF binding, it is currently more straightforward to study this potential from the angle of evolutionary TFBS data. We found that conserved TFBS data, compared to aggregate experimental TFBS data, are typically better suited for studying general patterns of transcription regulation, such as prioritizing which TFs may regulate a biological system (e.g. our studies of OXPHOS regulation in Chapter 6). Another way to infer cell type- and condition-specific TF binding is by combining information on DNA accessibility with motif searches [29]. For instance, genomic regions corresponding to DNase I hypersensitive sites under a particular condition can be scanned for known motifs in order to predict TFs that bind to these open chromatin genomic regions under that condition.

While the overall reliability of conservation-based TFBS predictions improves dramatically when more sequence data is available, a downside of the conserved TFBS data is that individual motif instances tend to still have a relatively high false positive rate [30]. Thus one should not put
a lot of trust in an individual binding site unless its existence can be verified, at the minimum through inspection of the alignment quality. At the same time, individual ChIP-seq peaks can also be unreliable. In fact, disagreement over the question to what extent the total combination of observed transcription, histone modification, open chromatin, DNA methylation or TF binding represents ‘function’, led to a heated scientific debate about the implications of the large, 80% fraction of the human genome reported to have ‘biochemical function’ by the ENCODE consortium [29,123,124]. Much of the discussion boils down to a decision to prefer reporting very many biochemical events (high sensitivity) at the cost of including many potential false positives (low specificity). Taken together, experimental and evolutionary approaches to characterize TF binding sites are complementary towards increasing our understanding of gene regulation.

Predicting regulators of pathways and systems from shared TF binding sites

TFBS data are instrumental in deciphering the transcription regulatory networks [125] underlying biological processes, development, and disease phenotypes. In addition to providing general insights into gene expression regulation, these networks contain information about which genes are involved in a molecular system on the basis of shared regulatory patterns. Genes sharing binding sites for the same TF, or a set of TFs, are likely to be co-regulated and may therefore be inferred to function in the same process. Vice versa, systematic assessment of common TF binding sites present in a set of genes that are known to make up a system, is a powerful approach for identifying which TFs regulate that system. There are many methods that exploit these concepts, though they differ widely in their exact approach and use different sources of TFBS data [126-132]. Some methods are quantitative and attempt to identify common motifs in genes that are co-expressed based on custom expression data, while others are more qualitative and for instance analyze a set of genes independent of their expression levels. Furthermore, some methods define the motifs regulating a gene based on a pre-calculated data set of TF binding sites, which can be in the form of presence/absence or have continuous confidence scores based on for example ChIP-seq peak strengths or conservation levels. Other methods use de novo identification of over-represented sequence motifs occurring in the genes of interest. For example, the ISMARA method requires user-provided transcriptomics data to identify those TFs that best explain the gene expression levels of their target genes [126]. It bases these calculations on a data set of conserved (across seven mammals) motifs in gene promoters, which together with the gene expression data are fitted to a linear model of TF ‘activities’.

To predict key regulators of biological pathways and systems, we ourselves developed the Conserved Motif Enrichment Method, CMEM (described in Chapters 2, 6 and 7). CMEM performs a statistical enrichment analysis for conserved transcription factor binding sites across a gene set of interest (see Section ‘Bioinformatics enrichment techniques’). For most applications we used conserved TF motifs putatively regulating human genes by processing the data from the 29 mammals project [30](see above). The CMEM method tends to retrieve the known regulators of a variety of molecular systems, and has therefore proved valuable for predicting novel regulators. In Chapter 2, we employed CMEM in studying the regulation of innate antiviral immunity, identifying IRF and NFκB motifs as statistically enriched in RLR pathway genes and using the known antiviral TF motifs IRF, NFκB, STAT, and AP-1 to predict novel antiviral genes. In Chapter 6, clustering, co-expression, and CMEM analysis of RNA-seq data from patients with OXPHOS complex I deficiency
revealed most well-established factors involved in OXPHOS biogenesis (including GABP, NRF1/2, SP1, YY1, E-box factors), as well as a set of novel candidates (ELK1, KLF7, SP4, EHF, ZNF143, TEL2). In Chapter 7, ciliary genes are shown to be strongly enriched for conserved binding sites of the RFX transcription factor family ($P = 10^{-11}$), as well as Pou4f3 and Hsf motifs. We made use of the well-established importance of RFX and FOXJ1 for cilium development to predict novel ciliary genes based on the presence of conserved motifs for RFX (x-box motif) and FOXJ1. In studies of intellectual disability (ID) genes, our CMEM analyses identified (i) several enriched TFs such as SOX genes and MEF2C that were known ID genes themselves, (ii) two enriched regulators, YY1 and SP1, mutations in which were later discovered to cause ID, and (iii) various candidate regulators of the ID phenotype, using both human data and neuron RNAi data from the fly model Drosophila melanogaster (the latter analysis identified CrebA) (RvdL, unpublished results, http://sysid.cmbi.umcn.nl/table/transcription-factor).

**Predicting miRNA-based gene expression regulation**

Similar to the DNA level, insights into post-transcriptional gene expression regulation at the RNA level are obtained from both experimental and computational approaches. For example, CLIP-seq experiments allow the large-scale characterization of miRNA-mRNA interactions. Theoretical miRNA binding sites in mRNAs can be predicted using knowledge about the seed sequences of miRNAs in combination with other requirements such as sequence conservation [133]. One of ten features exploited for finding antiviral RLR pathway genes in Chapter 2 uses an aggregate score of the predicted target sites of viral miRNAs in human gene transcripts.

**Protein interactions: within-species and between-species (host-pathogen) interactomes**

Most high-throughput characterization of physical interactions between proteins has been performed using yeast two-hybrid (Y2H) screens or some form of affinity purification followed by mass spectrometry (AP-MS) [134]. The Y2H technique is able to detect a binary interaction between two proteins if that interaction restores a functional transcription factor, which then activates expression of a reporter gene. This involves the non-physiological situation of cloning and over-expressing two proteins into a yeast strain and depends on them interacting in the nucleus. AP-MS depends on mass spectrometry of samples that are purified to only contain proteins that bind to a protein of interest (e.g., using antibodies), which enables characterization of multi-protein complexes. The more recent complexome profiling technique is able to separate and characterize native protein complexes from cell extracts and can identify unknown complexes and components [135,136].

**Human protein interaction networks**

Human protein-protein interactions (PPIs) and protein complexes are being charted extensively [137-139]. This has led not only to increasingly reliable and higher-coverage reference maps of the global human protein interaction network [138], but also to the characterization of interactions involved in specific biological functions. An example is the reconstruction of an interaction network involving 58 regulators of type I interferon production, under various stimulations that mimic bacterial or viral infection [140]. The resulting network consists of 401 interactions between 58 known type I interferon regulators and 260 identified interacting proteins. The systematic construction of a compendium of genes involved in the cilium also makes extensive use of proteomics information.
In addition to various other types of omics information, we integrated data sets from Y2H and AP-MS experiments dedicated to finding interactors of known ciliary components.

Like most aspects of molecular biology, protein interactions are highly context-dependent and will vary with conditions, disease and cell type, which dictate among others the availability of (competing) interaction partners in the cell and their post-translational modification status. Condition-specific networks are increasingly constructed, for example by predicting which binding events from the global interaction network can occur based on external data such as gene expression [141,142] or by predicting the impact of disease mutations on the reference network [143].

The BioGRID database [31] collects reported protein interactions, both from low- and high-throughput experiments. It contains ~231,000 unique physical interactions between proteins encoded by almost all human genes (~20,000) extracted from ~24,000 publications (build 3.4.131, December 2015). The various public PPI databases focus their curation efforts for including PPIs in different areas; hence each contains a unique set of interactions [144]. Some databases include only interactions from either high- or low-throughput experiments. Other databases integrate PPIs from the various primary databases, but differ in the precise subset of databases used. While these discrepancies to some extent complicate the choice of PPI data, they also offer opportunities for selecting the best data for answering a given question. For instance, to identify novel components of a molecular system, PPI data can be screened for interactions involving the known system components (Chapter 2). This initial prioritization may tolerate some degree of unreliability and benefits from including as many interactions as possible. In contrast, detailed investigation of a small number of proteins, for example resulting from a larger initial screen, typically requires a stringent set of only the most confident PPIs. Such high-confidence interactions could be defined as those that have been validated by multiple, small-scale experiments, with evidence from independent publications, preferably measured in a context that is relevant to the biological question addressed [142].

**Human-virus protein interactions: insights and data**

In addition to within-species interactions, there is a big interest in identifying interactions between proteins from different species, especially in host-pathogen relationships. Viruses depend on their hosts to replicate. Successful infection requires many interactions between virus and host proteins, including proteins involved in antiviral immunity [70]. High-throughput identification of human-virus PPIs has provided insights into the infection mechanisms of e.g. Epstein-Barr virus [145], Hepatitis C virus [146], Influenza [147], and HIV [148]. Other studies have taken a cross-virus approach to identify common and virus-specific infection strategies. Measurements of interactions between human proteins and 70 innate immune-modulating viral open reading frames (viORFs) from 30 viruses revealed 579 human proteins with a broad range of antiviral functions [149].

Computational analysis of host-pathogen PPIs collected from different studies revealed that many groups of pathogens target the same processes in the human cell, with viruses preferring to interfere with the cell cycle, apoptosis, nuclear transport, and the immune response [150]. Virus-interacting human proteins (VIHP) are specifically enriched for pattern recognition pathways, including the antiviral RLR pathway (Chapter 2), which suggests their common targeting by viral
antagonists. Integration of virus-host PPIs with the human interaction network demonstrated that VIHP tend be more central to the network and have more human interacting partners (i.e. are hubs) than do human proteins for which no viral interactions have yet been detected [146,149,150]. These studies suggest that the targeting of central players in the host network is an efficient strategy for viruses to influence host cellular processes. Analysis of virus-host protein interactions, both binary PPIs as well as structural models of protein complexes (e.g. between viruses and their host entry receptors), are also useful for prioritizing positively selected genes and amino acids to identify virus-host genetic conflicts and predict novel cases of viral interference (Chapter 5).

The PHISTO database collects host-pathogen PPIs from protein interaction databases [33]. It contains ~30,000 interactions between human proteins and ~350 virus strains (November 2015). Several thousand human proteins are currently reported to interact with at least one viral protein, with numbers ranging from ~2,500 (VirusMentha [151]), to ~2,600 (Chapter 2), to ~5,500 (PHISTO). As discussed above, the differences in these statistics stem from the different focus and inclusion criteria of the various PPI databases.

**Functional data from systematic genetic screens**

‘Breaking’ the cell is a useful, if crude, way to understand it. For instance, genetic perturbations by e.g. gene mutation, deletion or over-expression provide insights into genotype-phenotype relationships through the discovery of genes that contribute to a cellular function. Modern high-throughput techniques permit large-scale functional screening of genes whose perturbation affects a phenotype of interest. These screens can assay any cell-based phenotype that is accurate and amendable for large-scale measurements, such as detecting the expression of a reporter gene (e.g. luciferase intensity) and imaging cellular morphology (e.g. cell shape or the number of mitochondria) [152]. RNA interference (RNAi) screens have become widely used. They rely on the (partial) silencing (‘knock-down’) of gene expression though siRNA-based targeting of mRNAs [153]. More recently, the sensational development of the CRISPR-Cas9 system is providing new opportunities for sequence-specific DNA editing on a genome scale. CRISPR-based methods can achieve a variety of large-scale perturbations, including complete knock-out of genomic loci, repression, or activation of gene expression [154]. In addition to high-throughput screens aimed at providing initial candidate genes involved in a cellular function, functional genomics experiments such as RNAi are also a popular validation strategy for assessing the quality of computational predictions (see Section ‘Experimental validation’).

**RNAi screens of virus infection**

Like with protein interactions, RNAi screens have investigated properties of human cells by themselves as well as in the context of human-pathogen interactions. For example, a genome-wide siRNA screen identified 154 regulators required for NFκB activation when the pathway was activated by the Epstein-Barr DNA virus [155]. These NFκB regulator genes are strongly predictive for genes involved in detecting cytosolic RNA viruses (RLR pathway, Chapter 2). RNAi screening is also powerful for identifying host factors that either limit virus infection, as these may be part of the immune response, or factors that are required for optimal virus infection, as these represent potential antiviral drug targets. This concept has prompted large-scale RNAi studies into the infection mechanisms of a variety of viruses (reviewed in [12] and [13]), identifying several hundreds
of cellular factors required for infection by e.g. HIV [156] and Influenza [147,157]. Although these studies tend to focus on identifying potential therapeutic targets and are therefore mainly interested in factors required by viruses, they do also provide information about potential immune factors that limit infection (i.e. genes that increase infection rates when knocked-down). We identified 173 such antiviral host factors by aggregating data from seven large-scale RNAi studies of viral infection (Chapter 2), but found that these factors had limited predictive potential for components of the antiviral RLR pathway.

**Constructing data sets from functional genetic screens**

Systematic cataloguing of hits from functional genetic screens is no trivial task due to the variety in experimental designs, confirmation strategies, data analysis and statistical cutoffs used. The GenomeRNAi database contains data from 452 RNAi screens performed in human (v15, January 2016) [34]. There exists no comprehensive repository for functional screens assessing the host-pathogen interaction. Construction of a specific data set, such as antiviral host factors, from large-scale RNAi screens currently requires various manual steps, including literature searches for relevant virus-human RNAi screens, gaining a case-by-case understanding of the study design and statistical procedures employed, retrieval of the relevant supplementary data sets, and careful data processing to obtain the hits. Perhaps surprisingly, similar procedures are not uncommon for obtaining various kinds of custom genomics data sets.

**Genetics data and disease-associated variation**

In addition to reference genome sequences for entire species (see Section ‘Genome sequences’), scientists have access to a wealth of genetic data describing individuals and populations within species. Efforts such as the International HapMap and 1000 Genomes projects have described common patterns of human genetic variation from different populations across the world with increasing resolution. HapMap genotyped 1.6 million single nucleotide polymorphisms (SNPs) in ~1,200 individuals using SNP arrays [158], while the 1000 Genomes Project by 2015 completed whole-genome sequencing for ~2,500 genomes from 26 populations [35]. In addition, the recent sequencing of several extinct archaic hominins, such as the Neanderthals and Denisovans, has advanced studies into the origin and genetic composition of modern humans [59].

**Studying the genetic basis of disease through genetic associations and exome sequencing**

Human genetic variation and polymorphism data are invaluable for studying the molecular basis of phenotypes like diseases [5,159] and other traits (e.g. height [160]). Genetic association studies, for example in genome-wide form (GWAS), identify genetic variants that are statistically associated with a trait by comparing the genotypes of large groups of affected individuals and controls. This approaches is best suited for identifying a collection of common genetic variants contributing to a multifactorial trait [161]; e.g. single nucleotide polymorphisms (SNPs) occurring in over five percent of the population. Examples of such multifactorial traits include complex diseases, which are caused by multiple genetic – and environmental – factors, with each genetic variant typically having a modest effect [159]. The GWAS catalog collects published SNP-trait associations [162]. It contains >15,000 associations for ~1,500 traits (http://www.ebi.ac.uk/gwas/diagram, November 2015). In contrast to GWAS approaches for studying relatively ‘common’ diseases, the genetic causes of more rare Mendelian disorders can be studied through exome sequencing-based
characterization of rare genetic variations in a small numbers of affected individuals [163]. Chapter 3 describes the identification of a genomic region contributing to susceptibility to fungal infection through genetic association analysis of SNP data from 227 patients and 176 controls. In Chapter 4, we used various genomics data relevant to antiviral immunity to prioritize rare variants detected by exome sequencing of patients with recurrent herpes virus infections. The section on ‘Applications of integrative genomics’ goes deeper into the role of molecular data in prioritizing disease-causing variants.

**Genetic variation modulating the immune response to infections**

GWAS has been highly successful in identifying risk loci for infection- and autoimmunity-related diseases [5], for example through genotyping with the Immunochip SNP array [164]. In the innate immune system, genetic variants in pattern recognition receptors influences susceptibility to infectious disease [4]. For instance, variants in the *IFIH1* gene influence susceptibility to a variety of autoimmune diseases, type 1 diabetes, and interferonopathies [165]. Given the well-characterized role of *IFIH1* as a cytosolic receptor for viral RNA, we were surprised to discover a genetic and functional link between *IFIH1* and fungal infection (Chapter 3). A recent study produced a map of genetic variants that influence the expression response of dendritic cells to LPS, influenza virus and IFNβ between individuals (ImmVar cohort of 534 healthy people) [166]. The approach identified a variant that modulates expression of the IRF7 transcription factor, which influenced the response to influenza-induced type I IFN production. Another study measured genetic associations of 669 female twins to several immune cell traits such as immune cell abundances and surface protein expression levels [167]. The associations identified by these types of studies are informative for prioritizing causal variants associated with immune disease, although non-heritable influences also contribute substantially to immune system variation [168].

**Systematic descriptions of molecular systems through integrative genomics**

The wealth of genomics data described above provides great opportunities for the systematic discovery of novel components of any biological system. However, despite the obvious success of individual studies in the discovery of novel functions for single genes, or even sets of genes, discovering the complete compendium of genes involved in a molecular system is no trivial task. Indeed, it is likely that many functionally important genes and proteins involved in fundamental cellular processes and disease are unknown, given that new functions are constantly being uncovered, even for systems that are considered well-described (see Section ‘Applications of integrative genomics’). A major reason for this missing knowledge is that comprehensive, systems-level description of biological processes requires characterization of the cell from multiple molecular ‘dimensions’, including DNA, RNA, proteins, interactions, and regulation (Table 2) [11,12,14,15,17,18,169-172]. Thus, in addition to generating more (big) data sets experimentally, we should also try to utilize the full potential of existing data and go beyond analysis of the individual data sets. Integrative genomics represents a systems biology approach that aims to achieve understanding though combination of large-scale molecular data (Figures 1 and 2).

**Data integration: exploiting strengths and complementing weaknesses of individual data**

Individual genomics data tend to be strong at characterizing molecular systems from one perspective, but rarely present a comprehensive picture. Data types have specific characteristics
and systematic biases, and differ in for instance their sensitivity (ability to correctly identify positives), specificity (ability to limit false positives), coverage (what fraction of the total information, e.g. genes and proteins, does the data provide measurements for?), and variability (what is the variation in a data type across different experiments, cell types, organisms?). For example, RNA sequencing at high sequencing depth is considered high-coverage in the sense that it should be able to detect all except the rarest transcripts present in a cell, although sample preparation steps may introduce biases in detecting particular parts of transcripts [76]. In contrast, protein interaction data tend to have lower coverage and will only contain PPIs involving proteins that were specifically targeted by the experiment, although various solutions such as standardized protein tags have resulted in higher-coverage PPI networks [134]. In terms of variability, human genome sequences are currently estimated to be 99.5% the same between different individuals [173]. This figure is an order of magnitude higher than for example the biological variation in gene expression levels between different tissues [83]. Systematic data integration exploits the strengths of individual data sets, while the complementarity between different data tends to ensure that individual weaknesses become less prominent. Thus, integration has the potential to increase the overall signal-to-noise ratio in the data, thereby ensuring improved predictive ability.

**Statistics as a foundation for systematic discovery**

The identification of relevant data sources for a given target molecular system, the objective assessment of their usefulness, and the data integration steps are often not trivial [174]. A combination of strong bioinformatics and statistics is required to answer questions like: What significance should be given to evidence for functional associations provided by one data set, but not another? In other words, how should we objectively evaluate the ‘quality’ of data sets for identifying components of a system? What aspects should be taken into account when deciding which data sets to use? And how confident can we be that a gene is part of a system based on the total evidence presented across various data sets?

Simple approaches are sometimes sufficient to prioritize genes for a role in a system. One may determine the overlap of the top genes in all data, take the genes that are implicated by the majority of data sources, or filter for some set of specific requirements [10,16]. For example, to filter candidate genes linked to viral infection susceptibility, one may require genes to be expressed in immune cells, have PPIs with viral proteins, and show a certain level of genetic variation and divergence between populations and across species (a simplified version of our approach in Chapter 2). Although such overlap or filtering approaches are acceptable in some cases, they do not exploit the relative predictive strength of data sets nor do they produce confidence scores for the prioritized genes.

Statistical and machine learning strategies are instrumental to more systematic evaluation and integration of data. They generate (binary) classifications or continuous rankings of genes that inform on the likelihood of involvement in a molecular system. Popular classification algorithms used with genomics data include Support Vector Machines (SVM), Random Forests (RF), neural networks, and Bayesian classifiers. The theory and bioinformatics application of these methods have been extensively discussed [175,176]. Different classifiers differ widely in the way they combine data and in the transparency of their results. For example, SVMs can take into account non-linear
combinations of data [176]. Random forests [175] and Bayesian integration approaches provide transparent information on the importance of data variables for classification.

**Weighing evidence using probabilistic Bayesian integration**

Bayesian integration approaches are based on probability calculations and are therefore transparent in every aspect, which is probably a major factor contributing to their popularity [16,52,177-182]. We applied this technique in Chapters 2 and 7, where we integrated various heterogeneous data sets to predict genes involved in the innate antiviral response and the cilium, respectively. A detailed statistical description of the naive Bayesian integration methodology is provided in the Methods of Chapter 2. The following section will outline the key features.

Bayesian integration to describe a molecular system requires three main elements: (i) positive and negative gold standards, or training sets (i.e. confident sets of genes known [not] to be involved in a system), (ii) data sources that contain information about the system (e.g. genomics data such as gene expression, PPIs or co-evolution patterns), and (iii) a prior estimation of the total size (i.e. number of components) of the system. In step one, each data set is evaluated for its ability to predict components of the target system by calculating conditional probabilities: given a certain (range of) value(s) in the data, what is the chance to find a known positive and a known negative training gene? The conditional probabilities are insightful for direct comparison of the predictive power and quality of the data set towards inferring functional associations. Furthermore, since the probabilities are independent of the type of data and are robust against missing data, Bayesian approaches are well suited for integrating data of completely different, heterogeneous nature. In step two, genes will receive posterior probabilities reflecting the likelihood that they are part of the target system according to the total evidence presented across the data. These posteriors essentially represent the prior probability that a gene is part of the system, updated by the conditional probabilities as contributed by the individual data sources. In this way data are inherently weighed based on their predictive ability. Thus, ‘better’ data will contribute more to the predictions, while uninformative data are statistically excluded (Figure 1).

The probabilistic nature of gene scores produced by Bayesian integration approaches facilitates straightforward ranking of genes and informed decision-making on classification cutoffs. For instance, the probabilities can be interpreted as expectation rates for validation experiments and are therefore informative for selecting lists of predicted genes that are compatibility with the capacity (how many genes can be tested?) and design (what is the expected validation rate of the experiment?) of follow-up studies (see Section ‘Experimental validation’). Furthermore, although prior probabilities are formally required to calculate absolute posterior probabilities, the weighted integration of data using conditional probabilities always leads to an informative relative ranking of genes, even if reliable prior estimates are challenging or if these estimates turn out to be inaccurate. Finally, the modular nature of the Bayesian integration approach allows for straightforward ‘updating’ of the posterior probabilities through replacement or addition of novel evidence-presenting data sets.
Bioinformatics enrichment techniques for discovering properties of gene sets

Statistical over-representation – or enrichment – techniques are common in bioinformatics \(^{(183)}\). They can be used to interpret a set of genes by identifying common ‘properties’, such as function categories and signaling pathways. These statistical associations can be determined by asking whether the property occurs more often in a test set of genes compared to a background set of genes; for instance all genes in the genome or all genes for which a measurement was taken in the experiment. Significance calculations then estimate the probability that chance alone would have given rise to the observed difference in occurrences, which is captured in \(P\) values and enrichment scores (e.g. odds ratios of test versus background genes). Although enrichment analyses are relatively crude, their strength lies in the fact that they harness the power of numbers. This means that even though individual annotations are not always trustworthy, aggregation of data for many genes dramatically increases the signal and tends to lead to reliable predictions.

Perhaps the best-known application of enrichment analysis is to gain insights into gene ontology function terms associated with a gene set of interest, such as biological processes (immune system, cell cycle, etc.)\(^{(184)}\). In the same way, one can perform enrichment analyses based on any property \(^{(183)}\), to identify for example protein domains that are common in the components of a pathway (Chapter 2), over-represented signaling pathways among differentially expressed genes (Chapter 3), and transcription factors involved in regulating a biological system (Chapters 2, 6 and 7; see Section ‘Predicting regulators of pathways and systems’.

Applications of integrative genomics for discovering components of molecular systems

The number of publications taking an integrative genomics approach is increasing substantially (Figure 3), stimulated by the discussed innovations in the generation and analysis of large-scale molecular data. These and other studies have proven that integration of such information is powerful for discovering components, principally genes and proteins, of a wide range of human disease systems. We can distinguish four broad types of integration studies: (i) compendium studies that have attempted to construct the complete set of genes involved in a system, (ii) studies that have predicted the total set of functional interactions between genes and proteins in an organism, (iii) studies that have prioritized disease-causing genes and genetic variants using secondary data sources, and (iv) other studies.

Figure 3. The growing number of publications applying an integrative genomics approach. To obtain a representative subset of studies, we retrieved all 186 journal articles with a combination of the words integrative/integrated genomics/omics in their titles by querying PubMed (February 2016): (“integrative omics”[title] or “integrative genomics”[title] or “integrated omics”[title] or “integrated genomics”[title]) AND “journal article”[Filter]. The first of these papers was published in 2003 by Mootha et al.\(^{(185)}\). Bars represent the number of integrative genomics publications per year, normalized for the increasing total number of scientific publications. Absolute numbers of integrative genomics publications per year are inside the bars. The total number of scientific publications per year was obtained by querying PubMed: (1000:2050[dp]) AND “journal article”[Filter]. Total publications more than doubled from 2003 to 2015 (0.53 to 1.07 million); integrative genomics publications increased ~10-fold.
(i) Next-generation compendia of molecular system components

The organelles of the cell have been extensively investigated for their total protein content, with most large-scale studies employing single-data type approaches [186-188]. The MitoCarta inventory of mitochondrial proteins was the result of systematic integration of multiple heterogeneous data sources, including protein interaction data, co-expression calculations, large-scale microscopy of mitochondrial subcellular localization, and curated information from literature [53]. The original MitoCarta was primarily based on mouse data, but was shown to be effective for identifying human mitochondrial disease genes and has been extensively used [53,179]. The updated MitoCarta2.0 is based on improved versions of the original data [189], e.g. more sensitive homology mapping [190]. It consists of 1,158 human genes with an estimated 5% false discovery rate. These improvements represent a clear case of the usefulness of Bayesian updating (see Section ‘Bayesian probabilistic approaches’). In Chapter 7, we took a similar genome-wide approach to systematically catalog proteins involved in the cilium organelle through Bayesian integration of experimental data, including protein interactions and expression data, and computational data, including transcription regulation, co-evolution and co-expression calculations. The CiliaCarta compendium comprises 836 genes with a confirmed or likely function in the cilium. Extensive experimental work has validated our predictions of novel ciliary genes, leading to the discovery of novel cilium biology (e.g. OSCP1 [Chapter 7] and TMEM107 [51]).

Besides organelar protein content, the total set of components of numerous cellular pathways has also been subject to investigation. For example, combination of phylogenetic profiles and other large-scale data sources identified genes involved in the small RNA (miRNA and RNAi) pathways [52]. Integration of heterogeneous data has also led to the expansion of immunity pathways [12,14,15], including innate pathways involved in recognizing viral infection. Pathway proteins involved in recognizing cytosolic DNA and retroviruses were identified by combining proteomics, genomics and small molecule perturbation screens [191]. We ourselves discovered regulators of the RIG-I-like receptor (RLR) pathway involved in innate detection of RNA viruses (Chapter 2). In that study, we took a knowledge-based approach to identify predictive features for innate antiviral genes from publicly available genomics data. Bayesian integration of ten data sets led to a genome-wide ranking of RLR pathway genes, which was validated by RNAi experiments. Another recent study used an automated approach to score data sets from a large collection of genomics data for their relevance to a number of different immunity pathways, and then calculated per-gene likelihoods of involvement in those pathways (ImmuNet) [192].

(ii) Global functional interaction networks of genes and proteins

Another topic that has benefitted from data integration is the prediction of functional interactions. The presence of genes and proteins in the same topological module of a global cellular network may indicate a shared function, as such modules often correspond to molecular systems [25,181]. First successful in yeast, combination of protein interactions from different experiments, such as two-hybrid and pull-down screens, and other data that are only weakly predictive on their own [177] led to more accurate predictions of functional interactions describing more proteins than any individual data set [25,178]. Human functional interaction networks for a range of different biological systems were obtained by Bayesian integration of genomics data, statistically weighing those data sets that are most relevant to each specific process [180]. The popular STRING database predicts
functional interactions for a user-defined set of query genes through weighted integration of known PPIs, genomic context, conserved co-expression, and co-mentioning in literature [32,93].

(iii) Data-driven prioritization of candidate disease-causing variants and genes
GWAS and exome sequencing typically result in lists of genes and genetic variants associated with a disease (see Section ‘Genetics data’). As the mutations cannot all be characterized functionally, some form of prioritization is required to identify the most likely causal candidates. This can be achieved through interrogation with secondary functional data and biological knowledge [10,16,193]. Many gene prioritization methods rank candidate genes based on a similar behavior across genomics data to a set of genes known to be involved in a disease [10,194]. In Chapter 4, we investigated genes that harbor missense mutations in patients suffering from severe recurrent herpes virus (HSV-2) infections. We prioritized the variants identified through exome sequencing using information on antiviral pathways and function, protein-protein interaction networks, and gene expression regulation. This approach highlighted a variant in the ZBTB25 gene, which was then shown to correlate with the disease phenotype in one affected family: only affected members were heterozygous for the ZBTB25 missense variant, while non-affected members were not. In Chapter 3, initial hints coming from genetic association analysis that IFIHI is associated with Candida infection susceptibility required further confirmation. Expression and immunological data could rule out the involvement of three other genes that located to the same disease-associated genomic region and verified IFIHI as the causal gene. Thus, circumstantial evidence and secondary data support objective prioritization of disease genes and variants.

(iv) Other forms of genomics data combination
Extending genomics analyses with another layer of data often reveals previously shrouded biological patterns. Modern genomics studies tend to analyze at least two types of data. Although any combination of molecular data can potentially deliver insights, the data types most likely to be useful often follow naturally from the biological question that is being addressed. For instance, understanding the molecular states underlying different cells and conditions can be gained via gene expression measurements and subsequent identification of genes that show differential behavior (see Section ‘Gene expression’). Additional analysis of transcription factor binding sites regulating such differentially expressed genes may provide a deeper understanding of the underlying changes to the regulatory network (see Section ‘Predicting regulators of pathways and systems’ and Chapter 6). Often such analyses do not represent the sort of systematic, statistical data integration that is part of the compendium, pathway and prioritization studies described in the previous sections. Nevertheless it is worthwhile to appreciate the value of more ad hoc data integration.

While integration of heterogeneous data types inherently interrogates a system from different perspectives, integration of multiple instances of a single data type is also fruitful for obtaining higher-confidence, higher-coverage descriptions. Co-expression calculations based on thousands of individual gene expression experiments are powerful for detecting functionally related genes (see Section ‘Co-expression’). The same is true for integration of TF binding site data across many
cell types and conditions (see Section ‘Transcription factor binding sites’) and for correlating phylogenetic profiles across multiple genome sequences (see Section ‘Phylogenetic profiling’).

A final application of large-scale data integration that deserves mention pertains to the modeling of an entire cell in terms of its molecular processes and components [195]. This is now becoming marginally feasible as a result of large-scale characterization of the cell at multiple levels. The whole-cell computational model of *Mycoplasma genitalium* integrated information on the genome, transcriptome, proteome and metabolome and consisted of 28 sub-models describing a variety of cellular processes [196].

**Experimental validation to complement computational predictions**

Systems biology studies are most valuable when computational predictions are combined with experimental validation [10,11,17](Figure 1, Chapter 8). The purpose of validation in integrative genomics studies is at least twofold: (i) to evaluate the overall performance of the approach and determine the value of the resulting resource of predictions; (ii) to perform detailed follow-up of individual predictions and gain mechanistic insights in the biological system studied. To some extent, computational and statistical methods themselves can serve to validate predictions. Prediction performance can be assessed by statistical cross-validation, which involves making predictions based on only part of the training data and evaluating the performance using the data that was left out. Furthermore, interrogation by independent data sets is useful for testing the validity of the prediction results. Nevertheless, evidence obtained from specifically designed, independent validation experiments probably provides the best proof of success [197].

Acquiring suitable validation opportunities can be a major challenge. It requires excellent collaborations between the groups involved. Potential validation strategies should be an integral part of the early study design, so that the computational and experimental frameworks can be adapted to each other (Chapter 8). Several considerations play a role in designing validation strategies that best match the expected outcome of the computational framework [10]. Validation experiments should obviously test the biological system that is described by the predictions. The type of assay also determines the balance between its capacity and reliability; that is, how many predictions (e.g. genes associated with a system) can be tested and how accurate is the expected outcome. Examples include Chapter 2, where we developed a theoretical framework that integrated ten types of genomics data to obtain a resource of candidate genes involved in the innate antiviral response. To complement the various computational validation steps, our collaborators developed a medium-throughput siRNA screen to test 187 candidate genes for a function in viral RNA-mediated type I interferon production (50% validation rate). Similarly, the integrated CiliaCarta compendium predicted novel candidate ciliary genes, which showed a 67% confirmation rate (matching the estimated theoretical rate) for 36 genes tested across various experiments measuring cilium function in human, mouse, zebrafish and nematode (Chapter 7). Suitable validation experiments thus give further weight to computational analyses.
Outline of this thesis

This thesis presents a collection of six computational biology studies that are united by two principal aims: (i) to gain molecular insights into complex biological systems relevant to human disease by exploiting the challenging range of large-scale molecular data and (ii) to advance the methodology for applying and integrating such information. Our studies have focused on several different model systems, which fall into three broad groups based on their clinical relevance (Table 1): infectious disease (antiviral and antifungal immunity), energy deficiencies (disorders of the mitochondrion), and ciliopathies (disorders of the cilium). In all studies except for Chapter 5, our bioinformatics work has guided experimental work, leading to fruitful multidisciplinary collaborations with virologists, immunologists, geneticists and cell biologists.

The human immune system traditionally consists of an innate (inborn) arm and an adaptive (acquired) arm to counter infections by microbes such as bacteria, fungi and viruses. As part of the innate immune response, our genomes encode a set of cellular systems involved in pathogen recognition that initiate immune signaling upon infection. Chapter 2 presents the discovery of ~100 genes involved in the innate immune response against RNA virus infections; specifically the RIG-I-like receptor (RLR) pathway that contributes to type I interferon production. Guided by the biology and virology of the innate antiviral response, we systematically analyzed multiple layers of public large-scale genomics data, including gene expression, protein interactions, transcription regulation, genome sequences and functional screens, to identify ten properties that represent general signatures of the RLR system. Bayesian integration of these signatures resulted in genome-wide predictions for RLR pathway components, which we validated using both computational (by us) and experimental strategies (by our collaborators).

Genetic variation in pathogen recognition pathways can lead to autoimmune disorders as well as increased susceptibility to infections. In Chapter 3, we reveal an unexpected role for the MDA5 protein, encoded by theIFIH1 gene, in the immune response against fungal infection. So far, MDA5 had been well described as a receptor of viral RNA in the RLR pathway. In our study, the first hints for a function in antifungal immunity were obtained from transcriptomics data. We then performed a genetic association analysis and identified missense genetic variants inIFIH1 (and other genes) that predominantly occur in patients susceptible to blood infections with fungi of the Candida species. These genomics analyses were supported by further immunological studies by our collaborators. Together the data suggested that MDA5 plays a role in the antifungal immune response.

Chapter 4 describes an application of our resource of innate antiviral response genes from Chapter 2. Our collaborators had studied three patients suffering from severe and recurrent infections with herpes simplex virus (HSV-2). Immunological studies showed a defect in the antiviral immune response and subsequent exome sequencing identified several hundred candidate genetic variants. To prioritize the responsible disease variants, we performed systematic overlaps with gene sets involved in antiviral immunity, including our RLR pathway resource. This highlighted several potentially causal variants per patient. A variant in theZBTB25 gene was subsequently shown by
our collaborators to co-segregate with the disease across a family of affected and non-affected individuals.

Viruses encode diverse strategies for interfering with the human immune system. This leads to rapid evolution and continuous genetic conflicts between viral and human genes. Chapter 5 studies the phenomenon of positive selection in the human-virus interaction through comparative evolutionary analysis of nine whole-genome sequenced primates. We developed a stringent analysis pipeline for enabling high-confidence inference of positive selection, exploring the challenges associated with obtaining reliable information from screens for positive selection. Our study presents a new conservative, but reliable lower limit (3%) of genes evolving under positive selection in the primate lineage. Immunity genes showed the strongest signal, but we also found evidence for positive selection in genes related to reproduction and other interesting processes. Investigation of the positively selected genes using genomic data describing the virus-host interaction identified novel candidate human-virus genetic conflicts as well as positions relevant to the structural interaction between viruses and their human entry receptors.

Defects of the mitochondria, particularly of the oxidative phosphorylation (OXPHOS) system, result in energy deficiencies. To better understand the molecular basis of mitochondrial disorders, Chapter 6 investigates the transcription regulation of the OXPHOS system. Our collaborators performed RNA sequencing of OXPHOS complex I-deficient patient cells and healthy cells, and investigated the transcriptomes through clustering analysis and co-expression calculations. We then performed a detailed analysis of the co-expression data and explored the underlying transcription regulatory network. We drew two main conclusions: (i) structural subunits and assembly factors of the OXPHOS system show distinct transcription patterns, and (ii) we implicated several novel transcription factors in OXPHOS biogenesis that could play a role in complex I deficiency.

In Chapter 7, we apply the Bayesian integration framework from Chapter 2 to a unique collection of genomics data to construct a global compendium of genes involved in the cilium. The cilium is an essential organelle of the human cell involved in a wide range of genetic conditions. Our computational work included the generation and analysis of seven proteomics, expression, regulation and evolutionary data sets. We then evaluated these data for their predictive value towards finding known ciliary genes and performed a systematic integration that formed the basis for a comprehensive compendium of 836 ciliary genes: CiliaCarta. A range of independent experiments in a variety of model organisms by our collaborators validated our framework and characterized OSCP1 as a novel ciliary gene. This chapter summarizes our contributions towards the design of this study, the analysis of the data sets, and the integration approach.

Chapter 8 concludes this thesis with a summarizing discussion and places the research into a broader perspective. We will argue that the data and techniques used in our studies are applicable to a wide range of molecular systems, discuss the role for big data in progressing biology, give a view on ongoing developments, and describe the implications and applications of our discoveries.
Chapter 2

Integrative genomics-based discovery of novel regulators of the innate antiviral response

*Equal contribution
**ABSTRACT**

The RIG-I-like receptor (RLR) pathway is essential for detecting cytosolic viral RNA to trigger the production of type I interferons (IFNα/β) that initiate an innate antiviral response. Through systematic assessment of a wide variety of genomics data, we discovered 10 molecular signatures of known RLR pathway components that collectively predict novel members. We demonstrate that RLR pathway genes, among others, tend to evolve rapidly, interact with viral proteins, contain a limited set of protein domains, are regulated by specific transcription factors, and form a tightly connected interaction network. Using a Bayesian approach to integrate these signatures, we propose likely novel RLR regulators. RNAi knockdown experiments revealed a high prediction accuracy, identifying 94 genes among 187 candidates tested (~50%) that affected viral RNA-induced production of IFNβ. The discovered antiviral regulators may participate in a wide range of processes that highlight the complexity of antiviral defense (e.g., MAP3K11, CDK11B, PSMA3, TRIM14, HSPA9B, CDC37, NUP98, G3BP1), and include uncharacterized factors (DDX17, C6orf58, C16orf57, PKN2, SNW1). Our validated RLR pathway list (http://rlr.cmbi.umcn.nl/), obtained using a combination of integrative genomics and experiments, is a new resource for innate antiviral immunity research.

**AUTHOR SUMMARY**

Viruses pose a continuous threat to human health, even though our immune systems have evolved to neutralize invading viruses. As part of the innate immune system, the RIG-I-like receptors (RLRs) are essential for detecting viruses during infection. Recognition of viral RNA by the RLRs triggers an antiviral response that inhibits viral replication, protects uninfected cells, and attracts specialized immune cells. Better understanding of the innate antiviral response may reveal novel targets for antiviral therapeutics and vaccine development. However, that requires knowledge about which genes and proteins are involved. In the present study, we systematically investigated the wealth of available genomics data (including gene expression, protein interactions, transcription regulation and genome sequences) and discovered no less than 10 distinctive properties of genes known to be part of the antiviral RLR pathway. By combining these properties in a statistical framework, we predicted 187 novel RLR pathway components. Our validation experiments showed that ~50% of the predicted candidate genes have a significant effect on antiviral signaling. These results, together with independent computational and literature-based confirmation, demonstrated the validity of our combined bioinformatics and experimental approach. Our study expands the collection of known antiviral genes, opening up new avenues for research into innate antiviral immunity.

**INTRODUCTION**

Viruses are a major cause of human disease, as highlighted by the pandemics of influenza viruses, HIV-1, and the current outbreak of the Ebola virus. Pattern recognition receptors (PRR) are among the first molecules that detect viruses during infection. The RIG-I-like receptors (RLRs, one class...
of PRRs) are part of the RLR pathway, which forms a crucial innate antiviral defense system [2,198]. Two RLRs, RIG-I and MDA5, reside in the cytosol where they recognize non-self 5′-triphosphate RNA molecules with short double-stranded regions and long double-stranded RNAs (dsRNA), respectively [199]. Activation of the receptors triggers a complex signaling network, key steps of which are the activation of the mitochondrial adapter MAVS, subsequent recruitment of the TBK1 and IKK complexes, phosphorylation/activation of IRF3 and NFκB, and translocation of these transcription factors to the nucleus. These steps ultimately lead to the production of type I interferons (IFNα/β) and proinflammatory cytokines, which are crucial for establishing an antiviral state in infected as well as neighboring cells, and also modulate the adaptive immune response [200].

The importance of the RLR system is further demonstrated by the observation that viruses of all types employ strategies to interfere with its activation, often at multiple steps [70,201]. Better understanding of viral interaction with the pathway has resulted in novel targets for the development of antiviral therapeutics and attenuated live vaccines, for example viruses lacking functional RLR antagonists [6]. Furthermore, mutations in RIG-I, MDA5, MAVS and other RLR pathway components are associated not only with strong susceptibility to infections, but also IFN-associated autoimmune disorders [9,202,203].

Previous studies into virus-host interactions and the innate antiviral pathways have used genomics approaches, often generating large data sets describing physical or genetic interactions [13,140,191,204]. Other publications have taken a comparative approach based on model organisms [205] or used over-expression screening systems [1,206]. Together, these studies have identified numerous genes with antiviral activity, including members of the RLR pathway. However, it remains important to systematically assess the quality of individual data sets as such screens report distinct sets of genes, often with limited overlap between them. Combining the many available genomics data sets in a statistical framework potentially allows for a more systematic discovery and categorization of genes involved in the RLR pathway. Indeed, Bayesian integration of large-scale data that includes weighing individual datasets for their predictive potential has been successful in other cellular systems, for example identifying novel protein interactions [177], mitochondrial disease genes [179], and small RNA pathway genes [52].

In this work we systematically exploit the wealth of available (gen)omics data, including transcriptomics and proteomics data, genome sequences, protein domain information, and functional genomics, to discover descriptive molecular signatures of the RLR pathway system. Bayesian integration of these data, together with comprehensive computational and experimental validation, confidently identifies novel genes involved in antiviral RIG-I signaling.

**Results**

The RIG-I-like receptor (RLR) pathway is a highly interconnected and diverse molecular system. We investigated whether available genomics data contain sufficient signal to accurately describe RLR pathway components, and whether such data could be used to prioritize novel genes for a possible role in RLR signaling.
Ten molecular signatures of RLR pathway components in genomics data

To discover molecular signatures that distinguish RLR pathway components from other genes, we explored a wide variety of genome-scale data describing different aspects of the virology and biology of the pathway. Some of these data we used directly, while other data were used as the basis for further calculations (Table 1). We quantitatively assessed the predictive power of each data set using a literature-curated standard of 49 known RLR pathway components from InnateDB [207] (‘RLR genes’, Figure S1) and a set of 5,818 ‘non-RLR genes’ that are unlikely to be part of the pathway (i.e. genes with known functions not directly related to the innate antiviral response, such as development, housekeeping and neurological processes, see Methods). Below we describe 10 signatures for predicting novel RLR pathway components. The first five signatures are based on the relationship of RLR genes with viruses, whereas the second set of five signatures are based on properties of the RLR pathway itself.

Virus-based signatures

Positive selection in primates. Viruses evade recognition or interfere with the immune systems of their hosts to achieve successful infection [6]. This involves interactions between virus and host proteins, the interfaces of which are under constant pressure to change [69]. We collected data on recurrent positive selection in the primate lineage, based on maximum likelihood analysis of sequence alignments [63]. RLR pathway components, e.g. the mitochondrial signaling adapter MAVS [211] and transcription factor IRF7 [212], are enriched for rapidly evolving genes (9% of 49 RLR genes) compared to genes that are unlikely to be part of the pathway (5% of 5,818 non-RLR genes, 1.7-fold enrichment, non-significant P = 0.38, one-tailed Fisher’s exact test, Figure 1A, Tables 1 and S1).

Protein-protein interactions (PPI) with viruses. The next signature is based on the physical interactions between host RLR pathway components and viruses. Viral proteins often interact with many host proteins during their infection cycle, including those involved in antiviral defense [72,150]. Extraction of virus-human PPIs from specialized databases [33] revealed ~2,600 human proteins that are reported to interact with at least one viral protein (virus-interacting human proteins, Figure S2). These virus-interacting human proteins include the majority of RLR pathway components (35/49 = 71%), while they include a significantly smaller fraction of non-RLR genes (1,000/5,818 = 17%, 4.2-fold enrichment, P = 1.4 × 10⁻¹⁶, one-tailed Fisher’s exact test, Figure 1A and Table 1). Among the RLR genes, TRAF2 (4 PPIs), DDX3, MAPK9, and the NFκB subunit RELA (3 PPIs each) have reported interactions with the largest number of distinct virus species.

Viral miRNA target. Another mechanism that viruses use to interfere with the antiviral activity of human cells is down-regulation of gene expression by miRNAs [213]. We collected 128 miRNAs encoded by nine, mainly herpes DNA viruses (Table S2) [208], most of which have confirmed physiological relevance. Predicted target sites of these miRNAs to the 3’UTR of human transcripts were then used to calculate for each gene a score representing the likelihood that viruses affect its expression (viral miRNA targeting score). For example, our method assigned IKBKE (IKKe) a relatively strong viral miRNA targeting score of 2.7. Indeed, Kaposi’s sarcoma-associated herpesvirus miR-K12-11 has been shown to inhibit translation of IKKe transcripts, leading to suppression of interferon signaling [214]. Analysis of the viral miRNA targeting scores revealed that RLR genes tend
to have stronger scores than non-RLR genes ($P = 0.03$, one-tailed Mann-Whitney $U$ test). Although the statistical significance of this trend is only marginal, we included it as a molecular signature of RLR genes as it still provides a moderate enrichment over non-RLR genes (1.3-fold enrichment

Table 1. Ten molecular signatures from genomics data used for predicting novel RLR pathway components

<table>
<thead>
<tr>
<th>Group</th>
<th>Molecular signature</th>
<th>Data set description</th>
<th>Type</th>
<th>References</th>
<th>Number of genes</th>
<th>Likelihood ratio score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virus-based</td>
<td>Positive selection in primates</td>
<td>Rapidly evolving genes in the primates lineage, detected using maximum likelihood analysis of nucleotide alignments</td>
<td>d</td>
<td>[63]</td>
<td>926</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>PPI with viruses</td>
<td>Virus-interacting human proteins extracted from PPI databases</td>
<td>c</td>
<td>[33]</td>
<td>2,587</td>
<td>4.2</td>
</tr>
<tr>
<td></td>
<td>Viral miRNA target</td>
<td>Likelihood scores of targeting of human transcripts by viral miRNAs, based on predicted target sites</td>
<td>c</td>
<td>[208]</td>
<td>6,761</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td>Differential expression upon infection</td>
<td>Genes showing differential expression in lung epithelial cells infected with four respiratory viruses</td>
<td>c</td>
<td>see Methods</td>
<td>1,680</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td>Antiviral host factor</td>
<td>Meta analysis of genes with antiviral activity from seven RNAi screens studying a variety of viruses</td>
<td>c</td>
<td>see Table S3</td>
<td>173</td>
<td>2.1</td>
</tr>
<tr>
<td>Pathway-based</td>
<td>Co-expression with RLR pathway</td>
<td>Weighted co-expression with known RLR genes across &gt;450 human gene expression studies</td>
<td>c</td>
<td>[90]</td>
<td>4,149</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>RLR pathway protein domain</td>
<td>Proteins containing one of the 25 domains that are significantly enriched in known RLR proteins</td>
<td>c</td>
<td>[209]</td>
<td>711</td>
<td>8.9</td>
</tr>
<tr>
<td></td>
<td>Innate antiviral TF binding motifs</td>
<td>Genes with IRF, AP-1, NFκB, or STAT TF binding motifs in their promoters, based on conservation across 29 mammals</td>
<td>c</td>
<td>[30]</td>
<td>4,508</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td>NFκB activation mediator</td>
<td>Hits from a genome-wide siRNA screen of Epstein-Barr virus-induced NFκB activation</td>
<td>d</td>
<td>[155]</td>
<td>154</td>
<td>19.8</td>
</tr>
<tr>
<td></td>
<td>RLR pathway PPI</td>
<td>Proteins that interact with known RLR proteins, calculated from PPI data</td>
<td>c</td>
<td>[210]</td>
<td>1,750</td>
<td>4.3</td>
</tr>
</tbody>
</table>

Integration RLR score Bayesian integration of the 10 molecular signatures to predict novel RLR pathway components

| | | |
| Integration RLR score | Bayesian integration of the 10 molecular signatures to predict novel RLR pathway components |

Data used directly (d) or as basis for further calculation (c). Combination of all bins with positive likelihood ratio scores per feature, derived from Figure 1A. RLR genes versus non-RLR genes: P(D|RLR genes) / P(D|non-RLR genes), see Methods. Note that, to avoid circularity, the predictive ability of the co-expression, protein domain and RLR pathway PPI data sets was assessed using the set of TLR, CLR, NLR, cytDNA genes instead of the RLR genes (see Methods). See also Figure 1A and Table S1.
among genes with the strongest viral miRNA targeting scores, Table 1), and even weak features can substantially improve the predictions for novel RLR genes.

**Differential expression upon infection.** Next, we asked whether RLR pathway genes are differentially expressed upon virus infection. To answer this, we used in-house gene expression data of human lung epithelial cells (A549) exposed to four respiratory viruses (respiratory syncytial virus, human metapneumovirus, parainfluenza virus, or measles virus), for which gene expression was measured at 6, 12 and 24 hours after infection (see Methods). Analysis of the transcriptomes revealed that many RLR genes (31%) underwent substantial expression changes ($\log_2$ fold change $>0.5$) in cells infected with the respiratory viruses, compared to the uninfected cells. This compares to a much smaller fraction of non-RLR genes (9%, 3.5-fold enrichment, $P = 1.3 \times 10^{-5}$, one-tailed Fisher’s exact test, Figure 1A and Table 1). The differentially expressed RLR genes include well-known interferon-stimulated genes (ISGs) like *ISG15*, *DDX58* (RIG-I), *IRF7*, *IFIH1* (MDA5), and *TRIM25*, which are the top five RLR genes most induced by the respiratory viruses studied ($\log_2$ fold change $>1.5$ compared to uninfected cells, Figure S3). Thus, even though we expect many RLR genes to already be expressed before viral infection, their expression levels are reinforced in infected cells.

**Antiviral host factor.** RNAi screening potentially allows the identification of host factors that limit virus replication, such as genes involved in the innate antiviral response, although most studies focus on hits with the opposite effect (i.e. factors required by viruses for replication) [13]. We performed a meta analysis of hits from seven large-scale RNAi studies in human cells, identifying 173 genes with antiviral activity against HIV-1, influenza, hepatitis C (HCV), West Nile, or enterovirus infection (Table S3). In contrast to our expectation that RLR genes would be common among these antiviral host factors, this data set is one of the weaker predictors of RLR genes: the 173 antiviral host factors contain only a single gene (IRF3) that belongs to the set of 49 known RLR genes (~2%) compared to 56 of 5,818 non-RLR genes (~1%, 2.1-fold enrichment, non-significant $P = 0.38$, one-tailed Fisher’s exact test, Figure 1A and Table 1).

**Pathway-based signatures**

**Co-expression with RLR pathway.** To aid in finding novel RLR genes, we screened >450 human expression studies for genes that co-express with known RLR pathway components using a two-step approach [90]. First, we weighed individual expression data sets for their propensity to predict new RLR genes: experiments in which the whole group of known RLR genes show high co-expression with each other received a higher weight and contributed most to the calculations. In the second step, we calculated the co-expression of all genes with the RLR genes. As expected, RLR genes display significantly higher co-expression scores with each other than with the rest of the genome or with non-RLR genes ($P \approx 10^{-27}$ for both comparisons, one-tailed Mann-Whitney $U$ test, Figure S4). However, RLR genes also score higher than components of other PRR signaling pathways (Toll-like receptor [TLR], C-type lectin receptor [CLR], NOD-like receptor [NLR], and cytosolic dsDNA sensing [cytDNA] pathways; $P = 4.5 \times 10^{-13}$, one-tailed Mann-Whitney $U$ test). Cross-validation by leave-one-out analysis confirms that the weighted co-expression approach retrieves RLR genes more readily than other PRR pathway genes, or genes involved in other aspects
Figure 1. Bayesian integration of ten molecular signatures of RLR pathway components from genomics data. (A) Distributions of the 49 known RLR pathway components (RLR genes, green) and 5,818 genes unlikely to be part of the pathway (non-RLR genes, red) across the 10 molecular signature data sets we identified as predictive of the RLR system (see also Table 1). Data sets were binned into discrete intervals and fractions of (non-)RLR genes add up to one. Arrows indicate the behavior of RIG-I across the data. The top five signatures describe the relationship of RLR signaling with viruses; the bottom five describe properties of the pathway itself. (B) Boxplots of the genome-wide integrated RLR score (Bayesian posterior probability score). Genes were grouped into one of five classes: known RLR genes (green, see [A]), components of other PRR signaling pathways: (TLR, CLR, NLR, cytDNA; purple), genes functioning in other aspects of the innate immune response (Innate immunity; blue), and non-RLR genes (red, see [A]). The remaining genes are classified as ‘other’ (gray). (C) The 50 genes with the highest RLR scores. Representative RLR and other innate antiviral response genes are indicated. The pie chart shows the occurrences of the different gene classes in the top 354 RLR ranks. (D) Receiver operating characteristic (ROC) curve illustrating the performance of the integrated RLR score (solid black line) and the individual molecular signatures (black dots) for predicting known RLR versus non-RLR genes. Sensitivity and specificity were calculated at various score thresholds (for the RLR score), or at specific thresholds that include all bins with positive likelihood ratio scores (for the individual data sets; see (A)). The asterisk denotes the sensitivity and specificity corresponding to a false discovery rate (FDR) of 57% (top 354 genes). Note that, to avoid circularity, the predictive ability of the co-expression, protein domain and RLR pathway PPI data sets in (A) and (D) was assessed using the set of TLR, CLR, NLR, cytDNA genes instead of the RLR genes (see Methods).

**RLR pathway protein domain.** Analysis of RLR pathway protein sequences revealed the presence of 40 unique domains, 25 of which were significantly over-represented compared to the full
human proteome (Benjamini-Hochberg-corrected Fisher’s exact $P < 0.01$, Table S4). These include protein kinase domains (12-fold enrichment, $P = 1.1 \times 10^{-8}$; present in IKKα/β/ε, MAP kinases, TBK1), the TBK1/IKK binding domain (TANK, TBK1BP1, AZI2), caspase and death domains (CASP8/10, FADD), IRF domains (IRF3/7), and the DExD/H box RNA helicase domain (15-fold enrichment, $P = 2.7 \times 10^{-3}$; RIG-I, MDA5, LGP2). We then assessed the domain organizations of all human proteins and determined a set of 711 proteins containing one or more of the domains enriched in RLR components. These proteins are predictive for RLR components with an enrichment score of 8.9 (Table 1).

**Innate antiviral transcription factor (TF) binding motifs.** Signaling through the RLR pathway triggers the activation of key transcription factors (TFs) like IRF3, IRF7, AP-1 and NFκB. Activation of these TFs leads to the production of type I interferons and proinflammatory cytokines that eventually activate STAT1 and STAT2 [2]. STAT1 and STAT2 in turn stimulate transcription of interferon-stimulated genes (ISGs), which include many RLR pathway components. To further explore the transcription regulation of RLR pathway components, we analyzed their gene promoters for the presence of TF binding motifs that are highly conserved across the genomes of 29 placental mammals, such as primates, rodents and many farm animals [30]. Conserved IRF and NFκB motifs are highly abundant in the promoters of RLR genes (Fisher’s exact $P = 3.3 \times 10^{-3}$ and $P = 2.0 \times 10^{-4}$, respectively; Table S5), suggesting the pathway is partly self-regulating as has been observed for individual components. Interestingly, a conserved IRF motif was detected not only in the promoters of IRF7 itself and in all three RIG-I-like receptor family members (DDX58 [RIG-I], IFIH1 [MDA5], DHX58 [LGP2]), but also in TRIM25, ISG15, and CYLD; three factors controlling RIG-I signaling activation by regulating the level of K63 polyubiquitination. In order to predict novel RLR components, we searched for genes containing conserved IRF binding motifs (several motif variants, collectively recognized by IRF1-9), STAT binding motifs (several motif variants, collectively recognized by STAT1-6), AP-1 binding motifs, or NFκB binding motifs (Figure 1A). We found 3,558 genes across the human genome containing one of these motifs in their promoters. This large number partially stems from similarities in the DNA binding preferences of TFs that belong to the same family, but does not mean that all identified genes are regulated by the RLR pathway. For example, STAT motifs not only occur in the promoters of ISGs, but also in genes involved in cellular proliferation, differentiation and apoptosis. Nevertheless, genes containing one of the four conserved TF motifs already show a good predictive value for RLR pathway components (enrichment score of 2.2, Table S1). Genes with more than one motif are even more likely to be RLR genes: 789 genes contain two motifs (2.6-fold enrichment) and 161 genes contain three or all four motifs (2.4-fold enrichment).

**NFκB activation mediator.** Host factors that regulate NFκB activation often also affect the RLR pathway. Indeed, the 154 hits that were picked up in a genome-wide siRNA screen of Epstein-Barr virus-induced NFκB activation [155] include a much larger fraction of known RLR genes (6/49 = 12%) than non-RLR genes (36/5,818 < 1%, 20-fold enrichment, $P = 1.0 \times 10^{-6}$, one-tailed Fisher’s exact test, Figure 1A and Table 1). Thus, these 154 NFκB activation mediators are likely to contain novel RLR pathway components as well.

**RLR pathway PPI.** Finally, to find novel RLR genes we assessed the human protein interaction network connecting the RLR pathway. PPI databases [210] report 3,504 interactions between 1,750
unique proteins and 47 of the 49 RLR components, the only exceptions being DAK and NLRX1. Of the 47 RLR proteins with reported PPIs, 41 are involved in a total of 147 interactions within the pathway (i.e. between two pathway members). This network of RLR components has significantly more connections with each other than do random networks of the same size and interaction degree distribution (physical interaction enrichment score = 3.6, $P < 1.0 \times 10^{-6}$ [215]). Using the PPI data, we obtained for each human protein the number of interacting RLR pathway components (Figure 1A). We found a total of 1,397 proteins reported to interact with one or two RLR proteins. These proteins are predictive for RLR components with an enrichment score of 1.8 (Table S1). A further 221 proteins interact with three of four RLR proteins (6.3-fold enrichment) and 132 proteins interact with five or more RLR proteins (15.8-fold enrichment). TBK1 (18 interactions), TRAF2 and MAVS (both 16) top the list, supporting their roles as central players in the RLR system [140]. Thus, an increasing number of interactions with RLR proteins indicates a higher likelihood that a protein is part of the RLR pathway.

Bayesian integration of molecular signatures provides genome-wide probabilities for RLR pathway components

The RLR pathway components published thus far probably constitute only part of the total proteins with a function in this pathway. To prioritize novel high-confidence genes for a role in the RLR pathway, we integrated the 10 identified molecular signatures of RLR genes in a naive Bayesian classifier [177,179] (see Methods). This approach weighs data sets based on their predictive value (i.e. their ability to separate known positives and negatives; Figure 1A, Tables 1 and S1) so that ‘better’ data contribute more to the predictions. Each human gene received a posterior probability score (‘RLR score’) reflecting the likelihood that the gene is part of the RLR pathway based on its behavior in the collected genomics data. A score of zero indicates equal probabilities of a gene being an RLR versus a non-RLR gene. Table S6 presents the genome-wide ranking of RLR scores (also available at http://rlr.cmbi.umcn.nl/).

As expected, known RLR pathway components have the highest RLR scores (Figures 1B and S5). Two-thirds (32/49) of these rank within the first 150 genes. The top ranking genes are IRF7, RIG-I, IKKε, subunits of NFκB, TRADD, TRAF2, MDAS, and IKKγ (NEMO) (Figure 1C). Other examples of well-described RLR pathway components include IRF3 (rank 51), ISG15 (52), MAVS (102), and LGP2 (114). Genes that are unlikely to play a role in the pathway (the set of non-RLR genes) generally have very low RLR scores, although some of these received high scores as well (Figure 1B). This is not unexpected, as even though this large set of genes was selected from function annotations generally unrelated to the innate antiviral response, this does not preclude that individual genes (also) function in the RLR pathway.

To gain insight into what kind of genes are present among the RLR predictions, we examined their functions by pathway and gene ontology enrichment analyses. The top 354 genes with the highest RLR scores (corresponding to high-confidence predictions, see below) have strong links with other pathways of the innate immune response, such as TLR, NLR, interferon, and cytokine signaling (Figures S6 and S7). Antiviral defense functions are also among the most frequent and significant terms associated with the high-scoring genes (Figure S8 and Table S7). Other important biological processes include various apoptosis-related functions, cancer and cell cycle pathways,
and regulation of metabolic processes and protein localization. Furthermore, the top predictions include a wide range of protein families, notably proteasome subunits, ubiquitin(-like) conjugating enzymes, and genes involved in phosphatidylinositol signaling (which was recently shown to affect the type I IFN response [204]). Finally, 22% of the top predictions are induced in cells treated with interferons (i.e. they are interferon-stimulated genes, ISGs) and ~18% are part of the common host transcription response to pathogens (Table 2). Together, these observations indicate that our framework successfully predicts genes with a likely role in the innate antiviral response and suggests other cellular systems and functions required for this response.

Performance estimates and independent data establish the reliability of the RLR score
We further computationally assessed the reliability of the integrated RLR score by estimating the sensitivity, specificity and false discovery rate (FDR) of the predictions using the positive (RLR genes) and negative (non-RLR genes) standards. Integration of the data sets achieved better sensitivity and specificity than any of the individual data sets (Figure 1D), thereby enriching for RLR genes and depleting false positives (Figures S5, S9 and S10). At an RLR rank threshold of 354 (RLR score -1.10), the framework correctly predicts 78% of the known RLR genes with a specificity of 98.4% (Figure 1D). At this threshold, only ~57% of the novel predictions are estimated to be false (Figure S11, adjusted FDR to match the expected total number of genes involved in the RLR pathway, see Methods). This compares to a genome-wide false discovery rate (i.e. when predicting

<table>
<thead>
<tr>
<th>Data set</th>
<th>References</th>
<th>Number of genes in data set</th>
<th>Fraction (number) of data set genes in top 354 predictions</th>
<th>One-tailed Fisher’s exact P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interferon-stimulated genes (ISGs)</td>
<td>[1]</td>
<td>354</td>
<td>22.0% (78)</td>
<td>$1.2 \times 10^{-67}$</td>
</tr>
<tr>
<td>ISGs with validated antiviral activity</td>
<td>[1,217]</td>
<td>45</td>
<td>42.2% (19)</td>
<td>$4.9 \times 10^{-73}$</td>
</tr>
<tr>
<td>Common host transcription response to pathogens</td>
<td>[81]</td>
<td>496</td>
<td>17.7% (88)</td>
<td>$5.2 \times 10^{-68}$</td>
</tr>
<tr>
<td>Interactors of the type I IFN protein network during pattern recognition (HCIP) a</td>
<td>[140]</td>
<td>241</td>
<td>11.2% (27)</td>
<td>$1.2 \times 10^{-15}$</td>
</tr>
<tr>
<td>HCIP with confirmed effects on IFNβ expression and antiviral activity</td>
<td>[140]</td>
<td>22</td>
<td>22.7% (5)</td>
<td>$1.9 \times 10^{-3}$</td>
</tr>
<tr>
<td>Tripartite motif (TRIM) family genes</td>
<td>[206]</td>
<td>71</td>
<td>12.7% (9)</td>
<td>$1.6 \times 10^{-6}$</td>
</tr>
<tr>
<td>TRIMs that enhance RIG-I-induced activation of IFNβ, NFKb and ISRE promoters</td>
<td>[206]</td>
<td>34</td>
<td>14.7% (5)</td>
<td>$1.7 \times 10^{-4}$</td>
</tr>
<tr>
<td>Human interactors of innate immune-modulating viral ORFs b</td>
<td>[149]</td>
<td>569</td>
<td>6.7% (38)</td>
<td>$4.7 \times 10^{-14}$</td>
</tr>
<tr>
<td>Genes expressed in PBMCs stimulated with Candida (CRG)</td>
<td>[216]</td>
<td>89</td>
<td>43.8% (39)</td>
<td>$4.7 \times 10^{-47}$</td>
</tr>
<tr>
<td>CRG with altered expression in CMC patients</td>
<td>[216]</td>
<td>23</td>
<td>65.2% (15)</td>
<td>$2.6 \times 10^{-22}$</td>
</tr>
<tr>
<td>Type I IFN response mediators</td>
<td>[204]</td>
<td>226</td>
<td>4.0% (9)</td>
<td>$9.5 \times 10^{-3}$</td>
</tr>
</tbody>
</table>

a These PPIs were not part of the RLR interaction network used for the RLR predictions (i.e. for the ‘RLR pathway PPI’ signature).

b These interactions were not used to determine the virus-interacting human proteins used for the RLR predictions (i.e. for the ‘PPI with viruses’ signature).
genes randomly) of ~99%. Thus, the integrated RLR score increases the probability of correctly identifying novel RLR genes by a factor of 43 compared to random classification.

Because we used the same gene sets for calculating the RLR scores and estimating the performance of the resulting predictions (i.e. without systematic cross-validation), there exists a danger of circular reasoning. Therefore, we also carefully validated the quality of the results using various independent and external data sets. First, we examined the high RLR scores for genes that have a known function in innate immunity, but not in the RLR pathway, and therefore were not part of our training set. Components of other PRR signaling pathways (TLR, CLR, NLR, cytDNA) have lower scores than RLR genes, but much higher scores than the rest of the genome (Figure 1B). The same is true for genes functioning in other aspects of the innate immune response (Figure 1B). Of the 225 novel predictions (i.e. those genes that are not part of the training sets) in the top 354 (FDR of 57%, see above), 142 (~63%) are part of these innate immunity gene lists (Figure 1C). Thus, the majority of high-scoring genes with no known link to the RLR pathway in fact have a function in other PRR pathways or other parts of innate immunity, supporting the relevance of our predictions.

Second, we compared our predictions to six recent data sets that are relevant to the innate (antiviral) response but that were in no way part of the RLR score calculations. The overlap with the 354 top genes, excluding known RLR genes, is significantly larger than expected by chance for all these data sets (Table 2). For example, the top predictions include: (i) 19 of 45 (42%) interferon-stimulated genes with validated antiviral activity against e.g. HIV-1, HCV, yellow fever, West Nile or chikungunya virus [1], (ii) 27 proteins from a set of 241 (11%) that interact with the type I IFN protein network during pattern recognition, among which are five confirmed modulators of IFNB expression and antiviral activity [140], (iii) nine tripartite motif (TRIM) family genes, five of which enhance RIG-I-induced activation of IFNB, NFkB and ISRE (IFN-stimulated response element) promoters [206], and (iv) 38 human proteins interacting with innate immune-modulating viral open reading frames (viORFs) from 30 viruses [149]. (v) Furthermore, the type I IFN response has recently been proposed to play a role in antifungal immunity [165,216] and the top RLR predictions are strongly enriched for genes expressed in PBMCs stimulated with the fungal pathogen Candida albicans: almost half (39/89 = 44%) of these occur in our top predictions (P = 4.7 × 10^{-47}, one-tailed Fisher’s exact test, Table 2). (vi) Finally, the overlap between our predictions and a genome-wide screen for regulators of RIG-I-mediated IFNβ production is, at only nine, marginal but significant (9/226 genes = 4%, P = 9.5 × 10^{-3}, one-tailed Fisher’s exact test, Table 2) [204]. In summary, these diverse and independent experimental data support the validity of our integrated RLR score for predicting genes with a role in the innate antiviral response.

**RNAi validation screens confirm the high predictive value of the integrated RLR score**

To further determine the predictive power of our in silico predictions, we selected 187 candidate RLR genes for experimental validation (Tables S6 and S8). These include 127 high-confidence candidates from the top 354, which have not been previously linked to the RLR pathway, supplemented with 60 candidates we selected from the top 1000 predictions, mainly on the basis of limited functional characterization in general (Figure 2A). Importantly, candidates with a known
Figure 2. RNAi screens validate a role for the novel RLR candidates in RIG-I-mediated IFNβ induction. (A) Flow chart of the RNAi validation screens. 187 candidate RLR genes were screened for RIG-I pathway activity in three different RNAi screens. In screens 1 and 2, HeLa cells stably expressing an IFNβ promoter-controlled firefly luciferase (Fluc) reporter were stimulated with a 5’-ppp-containing RIG-I RNA ligand. The 57 hits (15 up, 42 down) with the largest effect on IFNβ induction upon siRNA knockdown in screen 1 (stringent Z-score <-2 or >2) were tested again in screen 2 with a different set of siRNAs. The 19 top hits from screen 2 were then picked for screen 3, which is similar to the first two screens except that it measures IFNβ mRNA levels using quantitative real-time qRT-PCR. (B) Correlation between the negative control-based robust Z-scores of RNAi screens 1 and 2. The 57 top hits with Z-scores <-2 or >2 in screen 1 were tested again in screen 2 (purple data points). N.T., non-transfected; SCR, scrambled. (C) Overview of the 19 novel RIG-I pathway genes with the...
role in RLR signaling, other branches of PRR pathways, or apoptosis were excluded as we were most interested in finding novel components of the RLR pathway.

For the selected candidates we performed a medium-throughput RNAi screen (RNAi screen 1) using HeLa cells stably expressing an IFNβ promoter-controlled firefly luciferase reporter (HeLa-IFNβ-Fluc). To activate the RLR pathway and induce Fluc reporter expression we used a known small 5’-ppp-containing RIG-I ligand [218]. This setup led to specific activation of RIG-I, as RIG-I or MAVS siRNA transfection, but not MDA5 or scrambled siRNAs, resulted in loss of reporter activity (Figures 2B, S12 and S13). All negative controls (non-transfected, scrambled and MDA5 siRNAs) scored within 1.25 median absolute deviations of the plate normalized IFNβ induction levels (Z-score cutoff <-1.25 or >1.25, Figure 2B). At this cutoff, siRNA knockdown of 94 candidates (50% of all candidates tested) affected RIG-I-mediated IFNβ induction (Figures 2A and S13A-D, Table S8). Among these, knockdown of 59 genes decreased RIG-I-mediated IFNβ induction (down-hits) and 35 genes increased IFNβ induction (up-hits). It is important to note that the experimental approach only activates the RIG-I branch of the RLR pathway and will not confirm predicted RLR candidates that regulate MDA5 activation and downstream signaling to MAVS. Thus, among the 93 non-confirmed candidates, there might still be novel regulators of the MDA5-mediated IFNβ induction pathway, which should be further investigated. Altogether, the integrated RLR score is clearly a strong and reliable predictor for novel regulators of the RIG-I pathway.

From the 94 confirmed hits, we picked the 57 top hits with the largest effect (stringent Z-score <-2 or >2) for a second RNAi screen using a different set of siRNAs (RNAi screen 2, Figure 2A). In this second RNAi screen, only a single up-hit (7% of 15 up-hits tested) showed a Z-score >1.25. Besides this hit, two negative control wells also had a Z-score >1.25 (Figures 2B and S13E-H), which suggests that the single confirmed up-hit might be unreliable. The poor reproducibility of the up-hits might be attributed to the screening approach. For instance, we used a large amount of 5’-ppp-containing RIG-I ligand (see Methods), leaving limited room for increased pathway activation. In contrast, the second RNAi screen confirmed 26 down-hits at Z-score <-1.25 (62% of the 42 down-hits tested). Of these, 19 genes (45% of tested down-hits) could be confirmed at a conservative Z-score <-2 (Figures 2A-C, Table S8). Taken together, the two RNAi screens, guided by the predicted RLR candidates, have substantiated the validity of our approach and have revealed potential novel regulators of the RIG-I receptor pathway.

**Figure 2 (continued)** Large effects on IFNβ induction in screens 1 and 2 (Z-score <-2 in both screens). Black data points correspond to genes whose knockdown also causes a reduction in IFNβ mRNA levels in screen 3. (D) RNAi screen 3. 13 of the 19 top hits from screens 1 and 2 also reduce RIG-I-mediated IFNβ mRNA production (black bars). Experiments were performed in triplicate (n=3). Bars (mean±SEM) display the fold induction of IFNβ mRNA (corrected for actin mRNA levels) compared to the mock-treated control. Statistical significance was assessed by one-way analysis of variance (ANOVA) followed by Dunnett’s multiple comparison test, comparing the values for each of the 19 test genes to the combined negative control conditions (scrambled and LGP2, red bars). ** P < 0.01; *** P < 0.001. (E) Correlation between the in silico integrated RLR score and the probability of experimental confirmation in RNAi screen 1. The dark purple line represents all 94 hits with Z-score <-1.25 or >1.25; the light purple line represents the top 57 hits with Z-score <-2 or >2. The 187 experimentally tested genes were ranked-ordered based on the RLR score and precision was calculated sequentially as the fraction of validated hits among all tested genes having a certain RLR score or higher.
To gain further understanding of how the 19 top hits affect RIG-I-mediated IFNβ promoter activation, another RNAi screen was performed (RNAi screen 3). In contrast to the first two screens, here we did not use the IFNβ promoter-controlled Fluc reporter translation as readout, but we measured IFNβ mRNA levels using quantitative real-time (qRT)-PCR. As expected, knockdown of RIG-I and MAVS abrogated 5’-pppRNA-induced IFNβ mRNA transcription, while MDA5 knockdown [218] and LGP2 knockdown, which regulates only the MDA5-mediated IFNβ mRNA transcription, had no effect (Figure 2D). Of the 19 top hits from the first two RNAi screens, 13 genes (68%) in this third screen again showed a reduction in RIG-I pathway activation. Nine of these showed a significant reduction (NUP98, TRIM14, C16orf57, PSMA3, G3BP1, DDX17, MAP3K11, SNW1, CDK11B; \( P < 0.01 \), one-way ANOVA with Dunnett’s post hoc test; Figures 2A and 2D), suggesting that these gene products play a so far uncharacterized role in the RIG-I signaling pathway upstream of IFNβ mRNA transcription.

In summary, using RNAi-based screening methods we validated more than 50% of the tested candidates. To further assess the predictive power of the in silico integrated RLR score, we ranked the experimentally tested genes based on their RLR score and sequentially calculated the fraction of hits (either considering all 94 hits from RNAi screen 1, or only the 57 top hits) among all tested genes having a certain RLR score or higher (Figure 2E). Higher RLR scores were experimentally confirmed more often, indicating that these indeed correspond to more confident predictions. Further analysis revealed that there is no molecular signature that solely explains the predictions of the validated hits; rather the integrated score of the 10 molecular signatures is important (Figure S14).

**Discussion**

Knowledge about the constituents of biological systems and pathways is an essential step towards understanding their function in health and disease. In this study, we showed that existing biological data can be exploited to successfully identify novel components of a key intracellular defense pathway; the antiviral RIG-I-like receptor (RLR) pathway. The RLR pathway is important for detecting viral infections, and its dysfunction can increase susceptibility to infections with viruses [202] and fungi [165], but is also associated with autoimmunity [9,203]. We systematically investigated a large variety of genome-scale data for their ability to predict RLR pathway components, covering most of the important (gen)omics data types such as protein-protein interactions, gene (co-)expression, genetic interactions from RNAi screens, comparative genome analysis, and transcription regulation. In these data, we found five virus-based and five pathway-based molecular signatures of RLR pathway components, providing insight into the determinants of antiviral signaling and type I interferon production. Bayesian integration of the signatures led to the genome-wide prioritization of novel RLR pathway components. We subsequently validated the predictions by comparing them with various independent data sets and experimentally confirmed more than 50% of 187 selected novel RLR candidates for a role in RIG-I-stimulated IFNβ induction. These results reiterate the potential of computational assessment and combination of available biological data as a complementary approach to studies generating novel large-scale data sets.
Identification of predictive signatures using a knowledge-based approach

To identify defining signatures of RLR genes in genomics data, we largely depended on current knowledge of the biology of the RLR system and its relationship with viruses. For example, since previous studies had shown that viral antagonism of specific RLR pathway components is prevalent [70,201], one of the first features we investigated, and indeed established, was that human-virus PPIs are a general theme for the RLR pathway as a whole. Similarly, guided by previous observations, we demonstrated that RLR genes conform to the tendency of immunity genes to evolve rapidly and commonly contain innate antiviral TF binding motifs, such as IRF and NFκB, in their promoters.

We also included several criteria that are effective for many different biological systems, but were specifically aimed at predicting novel RLR pathway genes in our case, such as the RLR co-expression calculations and RLR protein domain occurrences. We decided not to include associations based on text mining of published literature (e.g. co-mentioning of gene names in abstracts), because such approaches in our hands only enriched for genes already known to be involved in the RLR pathway and therefore compromised our ability to identify novel candidates. Finally, we settled on using a total of 10 molecular signatures that are relevant and predictive for the RLR system. Inclusion of additional data sets, generated for example by future experimental techniques, and substitution of existing data with novel and improved versions, will likely refine this data-driven definition of RLR genes over time and lead to updated Bayesian RLR probabilities that could further improve prediction accuracy.

A major challenge in our study arises from the fact that the RLR pathway is highly interconnected with other intracellular pathways, such as other innate PRR pathways (e.g. TLR and cytosolic DNA sensing), the stress response pathway, mitogen-activated protein kinase (MAPK) signaling cascades (e.g. TRAF2 and 6 lead to the p38 MAP kinases), and apoptosis (e.g. via CASP8 and 10) (Figure S1) [2,14,140,219]. Although our approach for predicting novel RLR components relied on a well-defined set of genes known to make up the core of RLR signaling, the overlap with other systems was a potential confounding factor. For example, most molecular signatures of RLR genes identified here, especially the virus-based properties such as PPIs with viruses, rapid evolution, and differential expression during infection, could also apply to genes involved in other aspects of antiviral immunity. Nevertheless, combination of the right signatures achieved reasonable specificity for RLR genes (Figure 1). Thus, we have extended an approach previously used for identifying components of membrane-enclosed organelles such as the mitochondrion [179] and showed that it is also possible to capture the complexity of a diverse and interconnected intracellular signaling pathway. The presented approach for identifying predictive signatures, followed by Bayesian integration, could potentially be applied to any cellular system.

Contributions of the individual signatures to the RLR predictions

Using the sets of known RLR and non-RLR genes, we could systematically assess the relative quality of the individual data sets for predicting novel RLR genes. Indeed the 10 molecular signatures have different predictive values as shown by the likelihood ratio scores (Figure 1A, Tables 1 and S1), and thus contribute with different weights to the integrated Bayesian RLR score. The data types with the strongest predictive value include NFκB activation mediators, RLR pathway protein domains, and both PPI signatures (PPIs within the RLR pathway and PPIs between human and viral
proteins). In contrast to our expectations, antiviral host factors identified in high-throughput RNAi experiments had a relatively small contribution. Besides raw predictive ability, we also considered the coverage of the data sets. For example, there are only few (<200) NFκB activation mediators and antiviral host factors, while the data sets on RLR co-expression, viral miRNA targets, and innate antiviral TF binding motifs identified many more genes (>4,000). Integration of all data sets with their varying coverage and predictive value into a single RLR score resulted in a classifier that is superior to the individual data sets (Figure 1D). This is underscored by the observation that the individual signatures by themselves are unable to explain the predictions for the experimentally validated RLR candidates, and only the integrated RLR score explains all validated genes (Figures 2E and S14).

Independent studies validate additional RLR candidates
Aside from our own experimental validation strategies, recent independent studies have confirmed a role for 15 of our predicted RLR candidates in the RLR pathway during viral infection (Table 3). Most of these publications appeared during the course of our study, and thus were not part of the knowledge or data used for predicting novel RLR genes. For example, TRIM14 (RLR rank 491) has been demonstrated to interact with MAVS leading to activation of IRF3 and NFκB via IKKγ (NEMO) [220]. Indeed, our predictions marked TRIM14 as a strong candidate RLR gene and all our RNAi screens confirmed it as a component required for optimal RIG-I signaling (Figure 2). Two additional high-confidence RLR predictions for which we validated an effect in all three RNAi screens have recently been validated externally as well: G3BP1 [221] and CDC37 [191].

Of the 15 genes recently described in the literature, 11 were part of the candidate RLR genes tested in our RNAi screens (Table 3). Of these, seven genes affected RIG-I-mediated IFNβ induction in RNAi screen 1 (Z-score <-1.25 or >1.25) and showed a consistent effect in RNAi screen 2. Therefore, our experimental screening condition appears to detect these described RIG-I pathway regulators with a sensitivity of ~64% (7/11). Furthermore, four out of four down-hits from our experiments (i.e. genes that decreased IFNβ induction when knocked down, hence positive regulators) that have been described in the literature were indeed described as positive regulators of RIG-I signaling (Table 3). Given that our experimental approach detected most, but not all, of the published RIG-I regulators, a substantial number of our predicted RLR candidates not validated by our RNAi screens might still play a role in for example a different cell type, downstream of type I IFN production, or regulate the pathway via MDAS/LGP2 activation. For example, RNF114 (RLR rank 181, Z-score RNAi screen 1 = 1.68, Table 3) is an ISG and therefore needs to be up-regulated via a positive feedback loop to fully contribute to RLR pathway stimulation [222]. This gene was not confirmed in all RNAi screens, perhaps because the time of RIG-I stimulation in our screens (6 hours) was simply too short. Similar biological reasons could limit the detection of an effect for other genes as well. Therefore we conclude that the hits identified in our RNAi validation experiments may be a conservative estimate of the number of correct RLR predictions.

Novel RIG-I pathway components DDX17 and SNW1 could regulate activation of transcription factors NFκB and IRF3
We identified 13 novel RIG-I pathway regulators that reduced IFNβ induction in all three RNAi screens (Figure 2). These include cell cycle gene CDK11B, heat shock protein HSPA9B, MAP kinase
Table 3. Validations of our predicted RLR candidates by independent studies

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Gene description</th>
<th>RLR rank</th>
<th>Described function</th>
<th>References</th>
<th>Type of regulation (literature)</th>
<th>Type of regulation (our RNAi screens)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSNK2A1</td>
<td>Casein kinase II subunit alpha</td>
<td>45</td>
<td>The casein kinase II complex inhibits the RIG-I-mediated antiviral response through phosphorylation of RIG-I</td>
<td>[223]</td>
<td>-</td>
<td>0&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>TRIM38</td>
<td>Tripartite motif-containing protein 38</td>
<td>56</td>
<td>Negative regulator of RIG-I-mediated IFNβ production by targeting AZI2 (NAP1) for degradation</td>
<td>[224]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RNF11</td>
<td>RING finger protein 11</td>
<td>78</td>
<td>Interacts with TBK1 and IKBKE (IKKε) to block TRAF3 interaction and restrict IRF3 activation</td>
<td>[225]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SMAD3</td>
<td>SMAD family member 3</td>
<td>100</td>
<td>Regulates dsRNA-induced transcriptional activation of IRF7 at the IFNβ promoter</td>
<td>[226]</td>
<td>+</td>
<td>0</td>
</tr>
<tr>
<td>UBE2D1</td>
<td>Ubiquitin-conjugating enzyme E2 D1</td>
<td>139</td>
<td>This Ubc5 E2 ligase is required for viral activation of IRF3 and MAVS by RIG-I</td>
<td>[227]</td>
<td>+</td>
<td>0</td>
</tr>
<tr>
<td>CDC37</td>
<td>Hsp90 co-chaperone Cdc37 (cell division cycle 37)</td>
<td>165</td>
<td>Regulates stability of TBK1 via Hsp90, allowing for induction of IFNβ in response to DNA viral and retroviral infections</td>
<td>[191]</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>RNF114</td>
<td>RING finger protein 114</td>
<td>181</td>
<td>Enhancer of dsRNA-induced production of type I IFN through positive feedback regulation</td>
<td>[222]</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>SRPK1</td>
<td>Serine/threonine-protein kinase SRPK1</td>
<td>235</td>
<td>Enhancer of RIG-I-dependent IFNβ and IFNα1 promoter activation during Sendai virus infection, possibly via IRF3/7 phosphorylation</td>
<td>[228]</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>CSNK2A2</td>
<td>Casein kinase II subunit alpha prime</td>
<td>249</td>
<td>The casein kinase II complex inhibits the RIG-I-mediated antiviral response through phosphorylation of RIG-I</td>
<td>[223]</td>
<td>-</td>
<td>0&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>G3BP1</td>
<td>GTPase-activating protein-binding protein 1</td>
<td>282</td>
<td>Functions in the formation of stress granules, which act as RLR signaling platforms that in some cases enhance IFN induction</td>
<td>[221]</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>UBE2I</td>
<td>SUMO-conjugating enzyme UBC9</td>
<td>284</td>
<td>Enhances RIG-I and MDAS SUMOylation, which correlates with increased IFNβ expression and repressed virus replication</td>
<td>[229,230]</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>SUMO1</td>
<td>Small ubiquitin-related modifier 1</td>
<td>326</td>
<td>IRF3/7 SUMOylation down-regulates IFN production; RIG-I/MDAS SUMOylation correlates with increased IFNβ expression</td>
<td>[229-231]</td>
<td>- / +</td>
<td>- / +</td>
</tr>
<tr>
<td>PPP1R15A</td>
<td>Protein phosphatase 1 regulatory subunit 15A</td>
<td>389</td>
<td>Required for IFNβ production induced by dsRNA and chikungunya virus in mouse; expression depends on PKR activation</td>
<td>[232]</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>TRIM14</td>
<td>Tripartite motif-containing protein 14</td>
<td>491</td>
<td>Interacts with MAVS upon viral infection, thereby recruiting IKKε (NEMO), which leads to activation of IRF3 and NFκB</td>
<td>[220]</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>DDX60</td>
<td>DEAD box protein 60</td>
<td>616</td>
<td>Promotes virus-induced, RLR-mediated type I IFN expression and increases binding of RIG-I to dsRNA</td>
<td>[233]</td>
<td>+</td>
<td>-</td>
</tr>
</tbody>
</table>

Total: 15 7 hits (out of 11)

<sup>a</sup> ‘+’: positive regulator (expected decrease in IFNβ induction upon knockdown). ‘-’: negative regulator (expected increase in IFNβ induction upon knockdown).

<sup>b</sup> Annotated cells (+, -, 0) indicate 11 candidate RLR genes that were tested in RNAi screen 1. ‘+’: down-hits from RNAi screen 1 (decreased RIG-I-mediated IFNβ induction upon knockdown, Z-score <-1.25). ‘-’: up-hits from RNAi screen 1 (increased RIG-I-mediated IFNβ induction upon knockdown, Z-score >1.25). ‘0’: no hit in RNAi screen 1, or inconsistent effect across RNAi screens 1 and 2 (CSNK2A1 and CSNK2A2, <sup>c</sup>).
MAP3K11, proteasome subunit PSMA3, nucleoporin NUP98 [234], and the recently identified RLR regulators CDC37 [191], G3BP1 [221] and TRIM14 [220] (Table 3). The remaining five genes, DDX17 (DEAD box helicase 17), C6orf58, C16orf57 (USB1, U6 snRNA biogenesis 1), PKN2 (serine/threonine protein kinase N2), and SNW1 (SNW domain containing 1), are overall least characterized. To obtain a first suggestion about how these genes might regulate RLR signaling, we searched for connections with the known human and viral protein interaction networks. Next, we discuss the reported interactions of DDX17 and SNW1 with the RLR pathway.

DEAD box RNA helicase DDX17 was recently found to bind Rift Valley fever virus RNA and restrict viral replication in an interferon-independent manner [235]. Our data now suggest a role for DDX17 in RIG-I-mediated IFNβ production as well. DDX17 has reported protein interactions with two other RIG-I regulators identified in our study: CDC37 and CSNK2A1. Interestingly, DDX17 also interacts with the peptidylprolyl cis/trans isomerase PIN1 [236], which inhibits RIG-I-mediated IFNβ production by inducing degradation of IRF3 [237]. Furthermore, DDX17 was present among a set of ISG15-modified (ISGylated) proteins in HeLa cells treated with IFNβ [238]. Thus, DDX17 could function in IRF3 activation by acting as a negative regulator of PIN1 and might be regulated by ISGylation (Figure 3). Lastly, DDX17 seems to be a preferred target of viral interference, having reported interactions with six different viruses (e.g. HIV-1 Rev and influenza virus A NS1, Figure 3).

SNW1 is an intrinsically disordered protein [239,240] that interacts with two other newly identified RLR regulators from our study, namely PKN2 [241] and C16orf57 [31]. SNW1 also interacts with the IKBKG (NEMO) protein [242], which is required for NFκB and IRF3 activation [243]. Given that our data shows that knockdown of SNW1 reduces IFNβ induction, SNW1 could be involved in NEMO

---

**Figure 3.** Human and viral protein interaction networks connecting the known RLR pathway with the newly identified RIG-I factors DDX17 and SNW1. Human proteins are represented by circles, viral proteins by rounded rectangles (purple nodes). Green nodes represent known components of the RLR pathway. Orange nodes (DDX17 and SNW1) are novel RIG-I pathway components discovered in our study, which are connected to the RLR network through interactions with the green nodes. Edges between human proteins represent physical interactions (both low- and high-throughput) obtained from BioGRID Release 3.3 [31]. Interactions between human and viral proteins were obtained from the PHISTO database (29 Sep. 2014) [33]. See Figure S1 for a more complete representation of the RLR pathway containing the curated set of 49 known RLR genes. LaCV, La Crosse virus; EBV, Epstein-Barr virus; SFSV, Sandfly fever Sicilian virus; PRRSV, Porcine reproductive and respiratory syndrome virus; HPV, Human papillomavirus.
regulation and thereby contribute to activation of the RLR pathway TFs, NFκB and IRF3 (Figure 3). The fact that SNW1 was also identified in a siRNA screen for mediators of virus-induced NFκB activation [155] strengthens this hypothesis. Further studies should be conducted to resolve the precise mode-of-action.

The genome-wide prioritization of RLR pathway components is a new resource for innate antiviral immunity research

We have validated the integrated RLR score with various experimental, literature and computational approaches. Our confirmations of a substantial fraction of the predicted RLR genes suggest the value of the prioritized list as a whole. The genome-wide prioritization of RLR pathway components is available in Table S6 and at http://rlr.cmbi.umcn.nl/, and can be used as a resource in several ways. For example, it can serve in the evaluation of data sets relevant to the innate antiviral and antifungal responses (Table 2). Many labs routinely consult internal data sets to decide which genes to study further. Comparison of such lists with for example high-scoring RLR candidates could provide insights into the quality of individual data sets for identifying antiviral genes and provide complementary hints about which genes could be important. Finally, the RLR resource could be used for prioritizing genetic variants in patients suffering from severe susceptibility to viral infections or inflammatory disorders caused by inappropriate production of type I interferons.

Conclusions

In this work, we have combined integrative genomics with experiments to discover 10 molecular signatures of a cellular signaling system that is central to human infectious disease: the innate antiviral RIG-I-like receptor (RLR) pathway. The described signatures span multiple layers of genomics data and provide new insights into the regulation of virus detection and immune signaling. Probabilistic integration of the data resulted in a confident genome-wide ranking of candidate RLR pathway genes. RNAi validation experiments confirmed 94 of 187 novel RLR candidates tested, including 13 novel factors with strong effects on antiviral signaling. These results, together with independent computational and literature-based confirmation, demonstrated the validity and high accuracy of our approach. Our study expands the collection of known antiviral genes, opening up new avenues for research into innate antiviral immunity.
METHODS

Human reference proteome and mapping
All data sets were calculated for and mapped to a reviewed reference set of 20,245 human proteins from UniProtKB/Swiss-Prot, release 2011_11 [244]. This set consists of one manually annotated record for each validated protein-coding gene. Gene/protein identifier mapping was performed using a mapping table from the same UniProt release. Ambiguously mapped identifiers were curated manually.

Molecular signature data sets
To systematically define RLR pathway components, we mined genome-scale data from a wide variety of sources. The data describe different aspects of the biology of the pathway; from the DNA to the protein level, highlighting evolutionary processes, virus-host interactions, sequence families, etc. We finally settled on 10 data sets that collectively distinguish RLR pathway components from other genes (see Table 1 for an overview and brief descriptions):

Virus-based signatures
Positive selection in primates. George et al. [63] calculated dN/dS-based likelihoods for recurrent positive selection across the exomes of seven primates (human, chimpanzee, orangutan, rhesus macaque, vervet, colobus monkey, tamarin). Maximum likelihood analysis of the nucleotide alignments of ~15,000 genes identified 930 genes with evidence for positive selection at \( P < 0.05 \). We grouped the genes according to these positive selection \( P \) values.

Protein-protein interactions (PPI) with viruses. Recent years have seen a surge of studies reporting interactions between viral and human proteins, both small- and large-scale. First, we collected all known virus-human PPIs from five specialized resources: PIG (9 Sep. 2011) [245], HPIDB (9 Sep. 2011) [246], VirHostNet 1.0 (24 Oct. 2011) [247], VirusMINT (6 Dec. 2011) [248], and PHISTO (25 Jan. 2012) [33]. These data were then combined to determine all human proteins for which an interaction was reported with at least one virus (Figure S2). Of note, the interactions reported by Pichlmair et al. [149] are not part of the final data set of virus-human PPIs and thus could be used for independent validation of the predictions (Table 2).

Viral miRNA target. Likely human target genes of viral miRNAs were determined in three steps. First, from the vHoT database we collected transcriptome-wide TargetScan (v5.0) predictions for the binding of 128 miRNAs from nine, mostly DNA viruses to the 3’UTRs of human mRNAs (Table S2) [208]. More negative scores are associated with more favorable binding site predictions. Second, because a single human transcript may be targeted (i) at multiple sites by a single miRNA and (ii) multiple times by different miRNAs, for each transcript we summed the prediction scores for all predicted target sites of all viral miRNAs. The resulting score (‘viral miRNA targeting score’) represents the overall likelihood that viruses target that mRNA. Third, the final score per gene was defined as the most negative score across its transcripts.

Differential expression upon infection. Human lung alveolar type II cells (A549) were cultured and exposed to four live respiratory viruses as described previously [249]: respiratory syncytial virus (RSV), human metapneumovirus (hMPV), parainfluenza virus type 3 (PIV), and measles virus.
(MV). RNA was isolated at 6, 12, and 24h post infection, as well as from mock-infected (medium without virus) control cells. Gene expression was then measured using the Affymetrix U133 plus 2.0 GeneChip platform and infection conditions were compared to uninfected cells. Data were log$_2$-transformed and normalized by VSN [250]. Statistically significant differential expression for each probe set (54,675 in total) was assessed using limma [251] and expressed as the fold change in expression between infected and uninfected conditions (FDR cutoff of 0.05). Genes represented on the microarray platform by multiple probe sets were summarized by the median differential expression across their probe sets. From the transcriptomics data we calculated for each gene (20,190 genes in total) the maximum absolute (i.e. considering both up- and down-regulation) change in expression across all time points and viruses, compared to uninfected cells. The column ‘Differential expression’ of Table S6 contains the processed gene expression data. A full analysis of these experiments will be described in a later publication.

220 and five genes were significantly up- and down-regulated in at least one infection condition respectively (>1.5 and <-1.5 log$_2$ fold expression changes). A total of 1,761 genes showed maximum absolute differential expression >0.5. Infection with hMPV induced maximal expression changes for the majority of genes (63% of the 1,761 genes with maximal absolute fold change >0.5), followed by RSV (29%). In comparison, PIV (4%) and MV (3%) caused less pronounced expression changes. Indeed, the expression profiles confirm these trends (Figure S3A). Furthermore, most genes tend to be increasingly up- or down-regulated during the course of infection (Figure S3A-B), with the distribution of expression changes becoming more extreme going from 6h (~5% of the 1,761 genes with maximal absolute fold change >0.5), to 12h (~14%), to 24h (~80%).

**Antiviral host factor.** We collected data from large-scale forward genetics screens aimed at identifying human genes involved in viral replication. Most studies focus on factors that reduce viral replication when inactivated, as these are often most abundant and represent candidate drug targets for infection treatment. However, these screens can also identify antiviral host factors, or host restriction factors, that inhibit virus replication (i.e. increase viral replication when inactivated). We collected the results from seven RNAi studies that investigated infection of human cells with a variety of viruses. These screens together reported a total of 173 unique antiviral host factors (Table S3).

**Pathway-based signatures**

**Co-expression with RLR pathway.** Functionally related genes tend to share expression patterns, i.e. be co-expressed. We employed an expression data integration method that weighs expression data sets for co-expression within a specific biological system [90]. From the NCBI gene expression omnibus database (GEO) [28] we obtained a collection of 465 publicly available human microarray data sets (~10,000 individual measurements). Each set of mRNA expression measurements was then assessed for its potential to find novel RIG-I-like receptor pathway genes by determining the coherence of expression of the 49 known RLR genes. That is, for each data set we ask whether known RLR genes behave similarly in terms of their expression, being up- or down-regulated together in the same microarray measurement. Sets of expression measurements in which known RLR genes show coherent expression receive a high weight, and will contribute more to the co-expression calculation than experiments with less coherent expression of known RLR genes. These
weights are then used to calculate an integrated score for each gene in the human genome, according to how much its expression profile correlates with that of the RLR genes across the expression data sets (Figure S4).

As the co-expression method was trained with the aim of retrieving RLR genes with high reliability, we also assessed its ability to retrieve RLR genes in leave-one-out cross-validation analysis. For that, we calculated the weighted co-expression 49 times, leaving out one of the 49 RLR genes in each fold (so that the whole set was left out exactly once), and determined the co-expression rank of the RLR gene that was left out. For the other gene sets (i.e. covering all genes except the RLR genes), we averaged the co-expression ranks across the 49 cross-validation runs. Figure S4D shows the recall (also known as sensitivity) of various gene sets at each rank cutoff in the cross-validation: genes were rank-ordered based on the RLR co-expression cross-validation rank and recall for each gene set was calculated sequentially as the fraction of genes among all genes in the set having a certain rank or higher.

**RLR pathway protein domain.** Domain organizations for all human proteins in SwissProt were obtained from the Pfam database (release 26.0; SwissPfam) [209]. We calculated statistical over-representation of domains occurring in the 49 known components of the RLR pathway compared to the background of all human proteins using the Fisher’s exact test. Enrichment P values were corrected for testing multiple domains (40 in total) using the Benjamini-Hochberg (BH) false discovery procedure and judged to be significant at a significance level of 1% (Table S4). Finally, we determined a set of proteins that contain one or more such enriched ‘RLR domains’.

**Innate antiviral transcription factor (TF) binding motifs.** Conserved TF binding sites in human were obtained from a comparative analysis of 29 genomes of placental mammals [30]. In this study, TF regulatory motif instances (putative TF binding sites) were detected across the human genome and assigned a likelihood based on conservation across the 29 mammals: for each motif match in human, the smallest phylogenetic subtree was calculated that contains the human motif and aligned motifs in other species [122]. To identify putative transcription regulators of a gene, we extracted conserved TF binding sites in promoter regions, which were defined as 4 kilobase (kb) windows centered (i.e. 2kb upstream and 2kb downstream) at all annotated transcription start sites of the gene [252]. We then searched for genes containing conserved motifs associated with four key innate antiviral transcription factors (IRF, AP-1, NFκB, and STAT; Table S5). Finally, we grouped all genes by the number of distinct motifs found: none, one, two, three or four.

**NFκB activation mediator.** Gewurz et al. undertook a genome-wide siRNA screen for NFκB pathway components [155]. They studied HEK293 cells with a stably integrated NFκB GFP reporter and inducible expression of Epstein-Barr virus latent membrane protein (LMP1), which activates NFκB. 155 LMP1 activation pathway components were identified, many of which are also important for IL-1β-, or TNFα-mediated NFκB activation. We obtained these hits and mapped them to 154 protein identifiers.

**RLR pathway PPI.** Human protein-protein interactions were obtained from the PINA database (release 28 Jun. 2011), which contained ~75,000 PPIs from six major resources [210]. We took all
interactions involving the 49 known RLR pathway proteins and counted how many interactions each protein is involved in, thus obtaining 1,750 proteins with at least one RLR interaction. Of note, the interactions reported by Li et al. [140] are not part of the final data set of RLR pathway PPIs and thus could be used for independent validation of the predictions (Table 2). We also assessed the cohesiveness of the RLR PPI network by calculating physical interaction enrichment scores, as described in [215].

**Training sets**

We assessed the capability of individual data sets to predict novel RLR genes using two ‘gold standard’ training sets.

**Positive gold standard.** We used a curated standard of 49 genes that are well characterized to play a role in the RLR pathway and make up its core (‘RLR genes’, all of which are depicted in Figure S1). This set is based mainly on the KEGG map [253] of the RLR pathway and taken from InnateDB (27 Mar. 2012) [207]. We focused on intracellular components, hence excluding the interferons and proinflammatory cytokines that are induced by the pathway.

**Negative gold standard.** To represent genes that are unlikely to play a role in RLR signaling, we constructed a set of 5,818 genes from seven functional categories generally unrelated to the innate antiviral response (‘non-RLR genes’, Table S6).

1. **Housekeeping genes.** These are typically defined as genes showing constitutive and constant expression in ‘all’ tissues. We collected housekeeping genes from five different studies [254-258], and included 1458 genes that were reported in at least three studies.
2. **Ribosomal subunits.** 134 human ribosomal (cytoplasmic and mitochondrial) proteins from [259].
3. **Transmembrane transporters.** 986 confirmed and predicted cytoplasmic membrane transporters and membrane channels from [260].

From QuickGO (6 Feb. 2012), we obtained human genes annotated with various gene ontology terms and their child terms [261], considering only annotations supported by experimental evidence codes (IMP, IGI, IPI, IDA, IEP, EXP):

4. **Mitoplast localization.** 559 genes with contributions to or co-localization with the mitochondrial matrix (GO:0005759) or mitochondrial inner membrane (GO:0005743). We did not include the inner membrane space and outer membrane, which is critical for RLR signal transduction through MAVS.
5. **Metabolism.** Genes with annotation ‘metabolic process’ (GO:0008152), excluding those annotated with the child term ‘protein phosphorylation’ (GO:0006468); 2243 genes.
6. **Neurological functions.** 1497 genes from GO term ‘neurological system process’ (GO:0050877).
7. **Embryonic development.** 775 genes from GO term ‘embryo development’ (GO:0009790).

We removed genes from the negative set that are known RLR genes, components of other PRR signaling pathways (TLR, CLR, NLR, cytDNA), or other innate immunity genes (see below). The resulting negative set is a good reflection of the rest of the genome in terms of the distributions
of the various molecular signatures and RLR integration scores (Figure 1B). Furthermore, given its size and the diversity of genes included, it is reasonable to expect a number of ‘non-RLR genes’ with high RLR scores. These should be considered as inappropriately included in the negative set and are therefore still candidate RLR genes.

**Other PRR pathway and other innate immunity gene sets**

Two additional curated sets of genes were used in our study (Table S6). The first consists of 153 genes with a known function (i.e. receptors, signaling components, etc.) in four PRR signaling pathways; the Toll-like receptor (TLR), C-type lectin receptor (CLR), NOD-like receptor (NLR), and cytosolic DNA sensing (cytDNA) pathways, but not in the RLR pathway. TLR, NLR and cytDNA components were obtained from InnateDB (27 Mar. 2012). We curated a list of 34 CLR pathway components, based mainly on [262]. The combined PRR pathway gene set was supplemented with several key proteins involved in virus-host interactions. As with the set of RLR genes, cytokines and other secreted proteins were excluded. The second list (‘other innate immunity genes’) consists of 803 genes with curated annotations from InnateDB (12 Jan. 2012) for a function in other aspects of the innate immune response, excluding RLR and other PRR signaling pathway genes.

**Naive Bayesian integration**

Individual (genomics) data sets contain important information about the make-up of cellular systems and pathways, but often have limited coverage and introduce data type-specific noise. Combination of multiple heterogeneous types of data, each approaching the characterization of a molecular system from a different angle, therefore has the potential to provide a more complete definition of the system and could have high power for predicting novel components involved.

We employed a naive Bayesian framework to facilitate direct comparison and weighing of many data sets describing properties of RIG-I-like receptor pathway components and integrate those data sets that were suitable into a single probabilistic score for each gene. Bayesian integration is well suited to combining evidence from dissimilar types of information and readily accommodates missing data [52,177,179]. Furthermore, this approach inherently weighs data sets based on their predictive value (i.e. their ability to separate known positives and negatives, Figure 1A and Table S1) so that better data contribute more to the predictions. Indeed, integration enriches for RLR genes and depletes false positive, non-RLR genes (Figures S5 and S9).

**Calculation of the RLR score**

For any given gene in the human genome, we can calculate the conditional probability that the gene is involved in the RLR pathway given the observed evidence in the 10 molecular signature data sets. More precisely, we calculated the posterior odds, defined as the ratio of the probability that the gene is an RLR gene versus the probability that the gene is not an RLR gene:

\[
O_{\text{posterior}} = \frac{P(\text{RLR gene}|D_1 \ldots D_{10})}{P(\text{non-RLR gene}|D_1 \ldots D_{10})}
\]

As this equation cannot be calculated directly, we approximate the ‘reverse’ likelihood ratio \( \lambda \) that a certain combination of values for the 10 data sets are observed, given the distribution of known RLR and non-RLR genes (i.e. the positive and negative training genes) across the data:
\[ L(D_1 \ldots D_{10}) = \frac{P(D_1 \ldots D_{10}|\text{RLR gene})}{P(D_1 \ldots D_{10}|\text{non-RLR gene})} \]

These two equations are related by Bayes’ theorem though the prior odds: the ratio of probabilities that any gene in the human genome is an RLR gene versus a non-RLR gene, prior to the use of information from our data sets. The prior odds can be calculated from the estimated total number of genes involved in the RLR pathway (see below).

\[
O_{\text{posterior}} = O_{\text{prior}} \cdot L(D_1 \ldots D_{10})
\]

\[
\frac{P(\text{RLR gene}|D_1 \ldots D_{10})}{P(\text{non-RLR gene}|D_1 \ldots D_{10})} = \frac{P(\text{RLR gene})}{P(\text{non-RLR gene})} \cdot \frac{P(D_1 \ldots D_{10}|\text{RLR gene}}{P(D_1 \ldots D_{10}|\text{non-RLR gene})}
\]

An assumption of the naive Bayesian approach is that the individual sources of evidence are independent of each other. Although this is rarely completely the case with genomics data, limited violations of the independence assumption still lead to effective predictions (see below). Under the independence assumption, \( L \) can be simplified and calculated as the product of the likelihood ratios of the individual data sets:

\[
L(D_1 \ldots D_{10}) = \prod_{i=1}^{10} \frac{P(D_i|\text{RLR gene})}{P(D_i|\text{non-RLR gene})}
\]

We calculated these likelihood ratio scores for individual data sets (Tables 1 and S1) directly from the contingency tables relating the positive and negative training genes to the data values binned into discrete intervals, asking: “What is the probably that a (non-)RLR gene has a value within a certain range in the data”? The bar plots in Figure 1A represent these contingency tables; likelihood ratio scores for each bin are defined as the ratios of the green versus red bars. Of note, as not all data sets contain values for all genes (e.g. genes can be missing from microarray platforms, were not tested in siRNA screens, etc.), we separated genes that were tested but show no effect from genes that were not tested. That is, we assigned no scores to bins that represent genes missing from the data entirely.

Having obtained the scores for the individual data sets and the prior odds, we then calculated the posterior odds that any gene is involved in the RLR pathway given its values in the data:

\[
\frac{P(\text{RLR gene}|D_1 \ldots D_{10})}{P(\text{non-RLR gene}|D_1 \ldots D_{10})} = \frac{P(\text{RLR gene})}{P(\text{non-RLR gene})} \cdot \prod_{i=1}^{10} \frac{P(D_i|\text{RLR gene})}{P(D_i|\text{non-RLR gene})}
\]

Finally, we obtained the ‘RLR score’ (Table S6 or http://rlr.cmbi.umcn.nl/) by \( \log_2 \) transformation of the individual terms in order to create an additive score:

\[
\text{RLR score} = \log_2 \left( \frac{P(\text{RLR gene}|D_1 \ldots D_{10})}{P(\text{non-RLR gene}|D_1 \ldots D_{10})} \right) = \log_2 \left( \frac{P(\text{RLR gene})}{P(\text{non-RLR gene})} \right) + \sum_{i=1}^{10} \log_2 \left( \frac{P(D_i|\text{RLR gene})}{P(D_i|\text{non-RLR gene})} \right)
\]

Taken together, the RLR score represents a Bayesian posterior probability, which depends on the positive and negative gold standard genes, the data sets used for the predictions, and the prior expected number of positive and negative genes in the genome. Although the RLR score may change for different priors (see below), the relative RLR ranks remain the same as these only depend
on the gold standards and the data. Thus, the relative ranking of genes as captured in the RLR rank is most informative.

**Conditional independence**
Although violations of the independence assumption can lead to over-estimation of the likelihood scores, previous work has shown naive integration of genomics data to be effective for predicting novel genes involved in a molecular system \([52,179]\). Assessment of the pairwise correlations between the 10 genomics data sets used for predicting RLR genes suggests that they are largely complementary (Figure S10). Several data sets have higher pairwise correlations, such as ‘PPI with viruses’ and ‘Innate antiviral TFs’. However, these features describe different molecular processes, namely protein-protein interactions between viral and human proteins and the presence of specific TF binding motifs, and hence can be considered largely independent in molecular terms.

**Performance estimates**
The performance of each of the 10 individual data types, as well as the integrated RLR score, for predicting RLR genes was evaluated using the positive and negative training sets. Based on these sets of known (non-)RLR genes, we calculated for each RLR score threshold (where genes with scores equal or higher than the threshold are predicted positives, i.e. predicted RLR genes, and genes with lower scores are predicted negatives, i.e. predicted non-RLR genes) the number of predictions that are:

- true positive (TP, number of positive training genes predicted as positive)
- false positive (FP, number of negative training genes predicted as positive)
- true negative (TN, number of negative training genes predicted as negative)
- false negative (FN, number of positive training genes predicted as negative)

These were then used to calculate several performance measures:

- **Sensitivity (SN)** \(\frac{TP}{TP + FN}\)
  
  fraction of positive training genes correctly predicted as positive (Figure 1D)

- **Specificity (SP)** \(\frac{TN}{FP + TN}\)
  
  fraction of negative training genes correctly predicted as negative (Figure 1D)

- **False Discovery Rate (FDR)** \(\frac{FP}{TP + FP}\)
  
  fraction of positive predictions that are false (i.e. that are negative training genes)

Calculation of the FDR depends on both the positive and negative gold standard genes. As the sizes of these training sets do not accurately reflect the expected numbers of RLR and non-RLR genes in the genome (prior probabilities, see below), we corrected the FDR to get an unbiased estimate using the following equation \([179]\) (Figure S11):

\[
FDR_{corrected} = \frac{1 - SP}{1 - SP + SN \cdot O_{prior}}
\]
Prior estimation of the number genes involved in the RLR pathway

Determination of the probability of finding a gene in the genome with a role in the RLR pathway, prior to the use of additional information, requires an estimation of the expected total number of RLR genes. We estimated this at 300; six times the number of currently known RLR genes in the positive training set. The prior odds then become ~1.5%:

\[
O_{prior} = \frac{P(\text{RLR gene})}{P(\text{non-RLR gene})} = \frac{P(\text{RLR gene})}{1 - P(\text{RLR gene})} = \frac{300}{20,245 - 300} = 0.015
\]

The prior odds influence the absolute RLR score and the corrected false discovery rate. Importantly, however, the overall ranking of genes does not depend on the estimated number of RLR genes. To assess the impact of the prior on the RLR score and false discovery rate, we re-calculated these measures using lower (75), medium (200) and upper (1000) bound estimates for the number of RLR genes (Table S9). These results suggest maximum and minimum FDRs of 84% and 28% at rank 354 (compared to an FDR of 57% when using a prior of 300).

Separate assessment of co-expression, protein domain, and RLR pathway PPI signatures

As described before, a positive gold standard of 49 known RLR pathway genes was used for calculating the likelihood scores for individual data sets. However, three molecular signatures (co-expression, protein domain and RLR pathway PPI) originate directly from calculations based on this same set of RLR genes. To avoid circularity, we assessed the performance (sensitivity, specificity) and likelihood ratio scores of these data sets using a different, independent positive training set: components of other PRR signaling pathways (TLR, CLR, NLR, cytDNA, see above). This approach prevented over-estimation of the predictive ability of these data sets and ensured that the likelihood scores of all molecular signatures are in the same range.

RNAi validation screens for RIG-I pathway activity

Cells and RIG-I ligand

HeLa-R19 cells stably expressing Firefly luciferase under control of the IFNβ (IFNB1) gene promoter were generated using the pIFNβ-Fluc-NeoR plasmid, which was kindly provided by Wendy Barclay (263). Single cell clones were selected under G418 selection, and a mixed population of two positive clones was used for the screens. Cells were maintained in DMEM supplemented with 10% FCS in a humidified incubator with 5% CO2. As RIG-I ligand, we used 5’-ppp cloverleaf (CL) derived from coxsackievirus B3 (CVB3), a 90 nt ssRNA carrying a 5’ triphosphate group, which was transcribed in vitro as described previously (218).

Protocol RNAi screen 1 – IFNB luciferase

In RNAi validation screen 1, we tested 187 candidate genes (Tables S6 and S8) that were predicted to play a role in the RLR signaling pathway by the computational framework. siRNAs (Dharmacon on-target plus Smartpool) were purchased internally from the Cell Screening Centre of the Utrecht University Medical Centre (CSC UMCU). Scrambled (SCR) and MDA5-targeting siRNAs were included as negative controls. Polo-like kinase 1 (PLK1)-targeting siRNAs were included as a positive control for cytotoxicity, while RIG-I-, and MAVS-targeting siRNAs were included as positive controls for RIG-I pathway activity. The RIG-I signaling pathway was activated by transfecting cells with the 5’-ppp-
containing CVB3 CL RNA. Activation levels were assessed by measuring IFNβ promoter-controlled luciferase reporter activity at 6 hr post transfection (Figure S12).

Screen 1 was performed in four technical replicates. Briefly, 0.5 pmole siRNAs (in 5 µl) was spotted per well. On the day of transfection 0.3 µl Lipofectamine RNAiMAX was diluted in 15 µl Opti-MEM and added to each well. Plates were rocked gently to mix the components and incubated at room temperature (RT) for 15 min. Then, 7,000 HeLa-IFNβ-Fluc cells (in 80 µl) were added to each well and plates were returned to a 37°C incubator. At 2 days post siRNA transfection, growth medium was discarded, replaced by 100 µl fresh medium and cells were then transfected with the RIG-I ligand. Briefly, 200 ng ligand and 0.8 µl Lipofectamine 2000 were separately diluted in 25 µl Opti-MEM, and incubated at RT for 5 min. These components were then mixed, incubated at RT for 20 min, and added to each well. At 6 hr post transfection, one replicate of each plate was fixed in 4% PFA and stained with DAPI. This replicate was later scanned at the CSC UMCU, and DAPI-positive nuclei were counted per well as an indication of cell viability upon siRNA transfections. The other three replicates were lysed in 30 µl 1x Passive Lysis Buffer (Promega) and allowed to freeze at -20°C. To measure luciferase activity, cell lysates were mixed by pipetting, and 15 µl from each well was transferred to a measurement plate, which was read using an automated plate reader using the following parameters: inject 40 µl firefly luciferase substrate (Promega), mix for 1 second, 1 second delay, measure for 10 seconds.

**Protocol RNAi screen 2 – IFNβ luciferase**

In RNAi validation screen 2, we tested the 57 top hits with the largest effects in screen 1 (stringent Z-score of <-2 or >2; Table S8) using a different set of siRNAs, separately assessing the 42 down-hits (siRNA knockdown of which resulted in down-regulation of RIG-I-mediated IFNβ induction) and 15 up-hits (siRNA knockdown of which resulted in up-regulation of RIG-I-mediated IFNβ induction). For 48 of the 57 genes tested, siRNAs (1 pool per gene) were purchased from SIGMA (esiRNAs human library) and used at 1 pmole per well during transfection. For the remaining 9 genes, for which esiRNA products were not available, Silencer Select siRNAs were purchased from Ambion, and three oligos per gene were pooled at 1:1:1 ratio and transfected at 0.5 pmole per well.

Screen 2 was performed in six technical replicates. The protocol was in principle the same as for RNAi screen 1, except that the MTT assay using Thiazolyl Blue Tetrazolium Bromide (SIGMA) (three replicates) was used to assess cell viability instead of DAPI staining. For the MTT assay, 60 µl 80 µg/ml MTT in medium was added to each well 1 hr prior to cell harvesting. The plates were incubated to 37°C for 1 hr. MTT-containing medium was removed, and reactions were quenched by adding 150 µl DMSO per well. The resulting mixture was measured at 570 nm using a plate reader.

**Statistical analysis of RNAi screens 1 and 2**

Raw Fluc intensities (Figures S13A and S13E) displayed limited variation between plates and were normalized using a negative control-based robust Z-score [264,265], which expresses each well as
the number of median absolute deviations (MAD) its intensity deviates from the median of the negative controls (non-transfected, scrambled and MDA5 siRNA wells) on the plate:

\[
\text{Robust Z-score}(x) = \frac{x - \text{median(negative controls)}}{\text{MAD(negative controls)}}
\]

Replicate plates (n=3) were then summarized by taking the median of the robust Z-scores of the well across the three plates (Figures S13B-C and S13F-G). We observed a clear difference in IFNβ induction levels between the positive (RIG-I and MAVS) and negative controls (mock treatment, scrambled and MDA5; Figures 2B and S13). Furthermore, significant correlation exists between screens 1 and 2 (correlation between Z-scores of all 57 genes tested in both screens, including the controls: Pearson r = 0.61, P = 8.6 \times 10^{-15}).

To reduce the potential for false-positive results, toxicity of the siRNA treatment was assessed by measuring nuclei counts (DAPI staining) in screen 1 (n=1) and cellular activity (MTT essay) in screen 2 (n=3). Both readouts were normalized per plate by calculating the percentage of the median of the negative controls (non-transfected and scrambled wells) and clearly separated negative from positive (PLK1) toxicity controls. Only a few siRNAs reduced cell numbers by over 50% in screen 1 (Figure S13D). However, knockdown of none of the 57 genes tested in screen 2 reduced cellular activity by more than 50%; only COPA showed slight toxicity (MTT level compared to negative controls is 53%; Figure S13H and Table S8). Thus, the observed effects of the siRNA knockdowns on IFNβ induction are largely independent of siRNA-induced reductions in cell numbers or cellular activity.

**Protocol RNAi screen 3 – IFNβ mRNA**

We assessed the 19 top hits (Figure 2C and Table S8) with the consistent largest effects in both RNAi screen 1 and 2 (5’-pppRNA-induced IFNβ induction in HeLa-IFNβ-Fluc reporter cells, stringent Z-score < -2), again for an effect on IFNβ (IFNB1) mRNA expression in an independent set of experiments. For 16 of these 19 genes, siRNAs (1 pool per gene) were purchased from SIGMA (esiRNAs human library). For the other 3 genes, for which esiRNA products were not available, Silencer Select siRNAs were purchased from Ambion, and three oligos per gene were pooled at 1:1:1 ratio. This RNAi screen 3 was performed in 24-well clusters, and performed in triplicate. Briefly, 5 pmole siRNAs were diluted in 50 μl Opti-MEM and incubated 5 min at RT. Next, 1 μl Lipofectamine RNAiMAX was added and incubated another 20 min at RT. Then, 25,000 HeLa-R19 cells (in 500 μl) were added to each well and plates were returned to a 37°C incubator. At 3 days post siRNA transfection, cells were transfected with the RIG-I ligand. Briefly, 200 ng ligand and 1 μl Lipofectamine 2000 were separately diluted in 50 μl Opti-MEM, and incubated at RT for 5 min. These components were then mixed, incubated at RT for 20 min, and added to each well. At 6 hr post transfection, total cellular RNA was isolated using the NucleoSpin RNA isolation kit (Macherey-Nagel) according to manufacturer’s instructions. Isolated RNA was used for reverse transcription using the TaqMan reverse transcription reagents kit (Applied Biosystems) with random hexamers primers (Invitrogen) according to manufacturer’s instructions. Quantitative analysis of IFNβ mRNA levels was performed using the LightCycler 480 (Roche) as described before [266].
Software and tools

Plots, statistics and other calculations were done using custom Perl and SQL scripts, and the R statistical package [267] with additional packages gplots [268], ROCR [269] and RNAither [270]. One-way ANOVA with Dunnett’s post hoc test was performed using GraphPad Prism (GraphPad Software).

Acknowledgments

We thank John van Dam and other members of the Huynen lab for stimulating discussions, and Pavel Čížek for assistance with the web page. RvdL, QF, ACA and MAH were supported by the Virgo consortium, funded by the Dutch government (FES0908), and by the Netherlands Genomics Initiative (050-060-452). QF and MAL were funded by personal grants from the Netherlands Organization for Scientific Research (NWO-017.006.043 and NWO-863.13.008, respectively). RS is supported by the Metakids Foundation. MGN was supported by an ERC Consolidator Grant (#310372). FJMWK was supported by a ECHO grant from the Netherlands Organization for Scientific Research (NWO-CW-700.59.007). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

Designed the study, analyzed the data, and wrote the manuscript: RvdL MAH. Planned and performed the large majority of (computational) analyses, data mining, data analysis, and manuscript preparation: RvdL. Designed, performed and interpreted the RNAi screening experiments, wrote the corresponding part of the manuscript and helped prepare the rest of the manuscript: QF MAL FJMWK. Calculated viral miRNA targeting scores and contributed to the integration approach: RtH. Calculated RLR co-expression and contributed to the integration approach: RS. Contributed to study design: MGN. Contributed the gene expression data and helped design the study: ACA. Supervised the study: MAH.
**Figure S1. Overview of the 49 RLR pathway components used as positive gold standard in our study (**RLR genes**).**

We focused on components that make up the intracellular core of the pathway, hence excluding the interferons and proinflammatory cytokines that are induced. The depicted network is based on the KEGG map of the RLR pathway [253]. Only key interactions are depicted. In reality, the pathway consists of a complex network of interactions [271].
Figure S2. Venn diagram showing the overlap between the five virus-human protein-protein interaction resources. Values represent the number of human proteins for which an interaction was reported with at least one virus. The union of the five databases (2,587 proteins) was used as a molecular signature (‘PPI with viruses’) for predicting novel RLR pathway components.

Figure S3. Time-course transcriptome analysis of A549 cells infected with four respiratory viruses. (A) Differential expression was calculated as the log2 fold change comparing each infection condition to mock-infected control cells. We calculated for each gene the maximum absolute (i.e. considering both up- and down-regulation) change in expression across all time points and viruses, compared to uninfected cells. This data was used as a molecular signature (‘Differential expression upon infection’) for predicting novel RLR pathway components. Colored lines represent the five RLR genes with the highest maximum absolute differential expression (represented by the colored dots) across all infection conditions. (B) Summary of the distributions of log2 fold changes across the infection conditions. Most genes tend to be increasingly up- or down-regulated during the course of the infection. Furthermore, RSV and hMPV generally induced much larger expression changes than PIV and MV (see Methods).
Figure S4. Analysis of the weighted co-expression calculations for the RLR pathway. (A) Distributions of weighted co-expression with the RLR pathway, binned into discrete intervals, for the ‘other PRR signaling pathways’ gene set (TLR, CLR, NLR, cytDNA; purple) and the set of non-RLR genes (red). Although the genome-wide RLR co-expression scores (x-axis in panels A-C) were calculated based on the set of known RLR genes, to avoid circularity we calculated the likelihood ratio scores (Tables 1 and S1) of this feature (‘Co-expression with RLR pathway’) using the independent set of TLR, CLR, NLR, cytDNA genes (see Methods). This is the same plot as in the co-expression panel in Figure 1A. (B) Kernel density estimates and (C) boxplots of RLR co-expression scores for the various gene sets. Density estimates were calculated using a Gaussian kernel with a smoothing bandwidth given by Silverman’s rule of thumb, and were normalized to 1. P values were calculated using the Mann-Whitney U test. (D) Recall performance (sensitivity) of the weighted co-expression method for retrieving a fraction of the 49 known RLR genes (y-axis) given an inclusion cut-off rank (x-axis), across a 49x leave-one-out cross-validation (green) (see Methods). The recall performance of the method for other sets of genes across the cross-validation ranks is also shown to demonstrate the ability of our method to retrieve RLR genes specifically compared to other PRR pathway genes (purple), or other innate immunity genes (blue).
Figure S5. Distributions of the integrated RLR score for the positive (RLR genes) and negative (non-RLR genes) training sets. Integration of the 10 molecular signature data sets into the Bayesian RLR score enriches for RLR genes and depletes non-RLR genes compared to the individual data sets (see also Figures 1A and S9).

Figure S6. KEGG [253] pathway enrichment analysis of the top 354 RLR predictions excluding known RLR genes. Purple bars indicate PRR signaling pathways other than the RLR pathway (TLR, CLR, NLR, cytDNA), blue bars indicate additional immunity-related pathways. Enrichment was determined using the functional annotation tool of the DAVID suite version 6.7 [272] with default settings and a false discovery rate (q-value) of 0.01. Background: all human genes. See also Table S7.

Figure S9. Visualization of how integration of the 10 molecular signatures enriches for RLR genes and depletes non-RLR genes. Rank plots showing the top 100 genes in (on the right) six of the individual molecular signature data sets and (on the left) in the integrated RLR score. Only the six continuous (i.e. non-binary) signatures are depicted, because ordering of genes within the two classes of the binary signatures would be arbitrary. See also Figures 1B and 1C.
Figure S7. ClueGO [273] enrichment analysis of REACTOME pathways [274] in the top 354 RLR predictions. Nodes represent significantly enriched REACTOME terms (Bonferroni step-down corrected P < 0.01, background: all human genes) and are grouped (as denoted by the connecting edges) based on overlapping gene lists (connectivity measure κ > 0.4). Groups of similar terms are represented by the most prominent term(s). See also Table S7.

Figure S8. ClueGO [273] enrichment analysis of Gene Ontology Biological Process terms in the top 354 RLR predictions. Nodes represent significantly enriched terms (Bonferroni step-down corrected P < 0.001, background: all human genes) and are grouped (as denoted by the connecting edges) based on overlapping gene lists (connectivity measure κ > 0.7). Groups of similar terms are represented by the most prominent term. For conciseness, clusters having less than four terms are not shown. See also Table S7.
The FDR was adjusted

Figure S10. Correlations between the ten molecular signatures used for predicting novel RLR pathway components. Heatmaps depict pairwise Spearman’s rank correlation coefficients between the values in the molecular signature data sets for positive gold standard RLR genes (A), and negative gold standard non-RLR genes (B).

Figure S11. Rank-order plot of the estimated false discovery rate (FDR) of the RLR predictions. The FDR was adjusted to match the expected total number of genes involved in the RLR pathway (see Methods). The inset shows the same plot, zoomed-in on the lower-left region, and indicates occurrences of RLR (green) and non-RLR (red) genes. RLR rank 354 corresponds to an estimated FDR of ~57%.

Figure S13 (next page). Analysis of RNAi screens 1 (A-D) and 2 (E-H) for validation of the candidate RLR genes. See also Figure 2. (A,E) Q-Q plots (left) of the raw luciferase intensities against the quantiles of a theoretical normal distribution (plotted by RNAither [270]). Linearity suggests that the raw data resemble a normal distribution. Boxplots (right) show the distributions of the raw luciferase intensities. (B,F) Q-Q plots and boxplots of the normalized data, summarized over the replicate plates. Raw luciferase intensities were normalized using a negative control-based robust Z-score and summarized across replicate plates by taking the median Z-score (see Methods). Note that the gray distributions in (A-B and E-F) include occurrences of RLR (green) and non-RLR (in red) genes. RLR rank 354 corresponds to an estimated FDR of ~57%.
Figure S13 (continued) ... activity (measured by MTT essay, right y-axis in (H)) distributions. Cell counts and MTT essay are presented as the percentage of the median of the negative controls (non-transfected and scrambled wells). No correlation exists between the effects of gene knockdown on the luciferase activity Z-score and cellular toxicity. All data points close to 0% cell counts or MTT are from the positive toxicity control PLK1.
Figure S12. Pilot experiments for RNAi validation screens of candidate RLR genes. (A) Our essay uses HeLa-IFNβ-Fluc cells stably expressing an IFNβ promoter-controlled firefly luciferase reporter. We knocked down candidate genes using different siRNAs, transfected cells with a known small 5’-ppp-containing RIG-I RNA ligand derived from coxsackievirus [218], and measured Fluc reporter expression and cell viability after 6 hours in three technical replicates. Pilot experiments for RNAi screen 1 (B,C) and RNAi screen 2 (D,E), which used a different set of siRNAs. (B,D) IFNβ-Fluc reporter activity after treatment of HeLa-IFNβ-Fluc cells with the 5’-ppp-containing RIG-I RNA ligand and various siRNAs. This setup led to specific activation of RIG-I, as RIG-I or MAVS siRNA transfection, but not MDA5 or scrambled siRNAs, resulted in loss of luciferase reporter activity. (C,E) Toxicity essays. Only the death-control PLK1 severely reduced nuclei numbers. See Methods for details.

Figure S14. No molecular signature solely explains the predictions of the experimentally validated hits. Distributions of the 187 candidate RLR genes selected for experimental validation, across the 10 molecular signature data sets we identified as predictive of the RLR system (see also Figure 1A). RLR candidates were grouped based on the results from RNAi screen 1: no hit (gray), all hits from RNAi screen 1 (94 hits with Z-score <-1.25 or >1.25, dark purple), and top hits from RNAi screen 1 (57 hits with Z-score <-2 or >2, purple) (see also Figure 2). Fractions of genes in the same group add up to one. ‘NA’ bins represent genes for which there was no data in the respective molecular signature (note that these bins did not receive a score in the Bayesian integration, see Methods).
Supplementary Tables

Available at the PLOS Computational Biology website (http://dx.doi.org/10.1371/journal.pcbi.1004553) and at http://www.cmbi.umcn.nl/~rvdlee/thesis/.

- **Table S1.** Likelihood scores for the 10 molecular signatures of RLR genes.
- **Table S2.** List of the 128 viral miRNAs for which we obtained predicted target sites in human mRNAs.
- **Table S3.** Meta analysis of antiviral host factors from published RNAi screens.
- **Table S4.** Enrichment analysis of protein domains occurring in RLR pathway components.
- **Table S5.** Enrichment analysis of conserved IRF, AP-1, NFκB, and STAT TF binding motifs in the promoters of RLR pathway genes.
- **Table S6.** Genome-wide prioritization of RLR pathway components based on the integrated RLR score.
- **Table S7.** Function enrichment analysis of the top 354 RLR predictions excluding known RLR genes.
- **Table S8.** Detailed results of the RNAi validation screens.
- **Table S9.** Impact of the prior on the RLR score and false discovery rate.
Chapter 3

The RIG-I-like helicase receptor MDA5 (IFIH1) is involved in the host defense against Candida infections

*European Journal of Clinical Microbiology & Infectious Diseases (2015), 34 (5): 963-674
Abstract

The induction of host defense against Candida species is initiated by recognition of the fungi by pattern recognition receptors and activation of downstream pathways that produce inflammatory mediators essential for infection clearance. In this study, we present complementary evidence based on transcriptome analysis, genetics, and immunological studies in knockout mice and humans that the cytosolic RIG-I-like receptor MDA5 (IFIH1) has an important role in the host defense against C. albicans. Firstly, IFIH1 expression in macrophages is specifically induced by invasive C. albicans hyphae, and patients suffering from chronic mucocutaneous candidiasis (CMC) express lower levels of MDA5 than healthy controls. Secondly, there is a strong association between missense variants in the IFIH1 gene (rs1990760 and rs3747517) and susceptibility to systemic Candida infections. Thirdly, cells from Mda5 knockout mice and human peripheral blood mononuclear cells (PBMCs) with different IFIH1 genotypes display an altered cytokine response to C. albicans. These data strongly suggest that MDA5 is involved in immune responses to Candida infection. As a receptor for viral RNA, MDA5 until now has been linked to antiviral host defense, but these novel studies show unexpected effects in antifungal immunity as well. Future studies are warranted to explore the potential of MDA5 as a novel target for immunotherapeutic strategies.

Key Points

- Candida albicans induces MDA5, a cytosolic receptor of viral RNA, in macrophages.
- Genetic variants in MDA5 are strongly associated with candidemia.
- MDA5 expression is reduced in CMC patients.
- Candida-induced cytokine profiles are affected by MDA5 genotype.

Author Summary

Candida species are common human fungal pathogens inhabiting the skin, mucosa, and gastrointestinal tract. Candida colonization does not normally cause disease in healthy individuals, but this changes when the immune system is compromised. For example, mouth and vaginal infections are relatively common, and the presence of Candida species in the blood can lead to invasions into deep tissues and organs, which is often lethal. While it is now clear that both environmental and genetic factors contribute to the causes and severity of Candida infections, improving treatment requires a deeper understanding of the molecular players and immunological effects. In this study, we have discovered that an intracellular receptor of double-stranded RNA (which is typically made by viruses), MDA5, plays an important role in the inflammatory response against Candida albicans infections in humans. We found that MDA5 is highly expressed in white blood cells stimulated with C. albicans. We show that patients with mutations in MDA5 are prone to C. albicans blood infections. This is likely caused by altered levels of inflammatory cytokines produced by white blood cells, which we show indeed depend on genetic variants in MDA5. Future efforts should shed light on the potential of MDA5 as target for treatment of diseases caused by Candida.
INTRODUCTION

*Candida* species are one of the most common human fungal pathogens. Oropharyngeal and vaginal *Candida* infections are often diagnosed in the population, while systemic candidiasis is the fourth most common form of bloodstream infections in the US with mortality rates reaching up to 40% [275-278]. In addition to the known risk factors (immunosuppressive medication, parenteral nutrition, prolonged intensive care hospitalization, etc.), recent studies have described several genetic risk factors that contribute to the cause and severity of systemic candidiasis [279,280]. Several monogenic disorders that result in primary immunodeficiencies increase the susceptibility to *Candida* infection, as demonstrated for mutations in *CARD9* and *STAT1* [281-283]. Common genetic variants, for example in pattern recognition receptors (e.g. *Dectin-1* and *TLR1*) and interleukins (e.g. IL-4, IL-10 and IL-12B), also increase the risk of infection by affecting *Candida* recognition and cytokine signaling [280,284].

However, more insight is needed to identify host defense pathways that are suitable targets for novel immunotherapeutic approaches. Recently, we described that the type I interferon (IFN) pathway plays a central role in host defense against *C. albicans* [216]. In the present work, we demonstrate that MDA5 (*IFIH1*), a RIG-I-like receptor, until now described as a receptor of viral RNA that induces a signaling pathway leading to the production of type I IFNs, is directly involved in the inflammatory response against *Candida* infections in humans. To this end, we present complementary evidence based on transcriptome analysis, genetics, and functional immunological studies in knockout mice and in healthy humans, as well as in patients suffering from systemic candidiasis or chronic mucocutaneous candidiasis (CMC). This is the first time that a receptor of the RIG-I-like helicase family has been shown to be involved in the antifungal immune response.

MATERIALS AND METHODS

Transcriptome analysis of *Candida*-stimulated macrophages

CD14+ monocytes derived from healthy volunteers were differentiated into macrophages using M-CSF for 7 days. Macrophages (2x10^5/well) were stimulated for 4 or 24h with either culture medium, a wild-type *C. albicans* strain (UC820) or an *HGC1* null mutant strain. This mutant is unable to form hyphae and is therefore locked in a yeast form, although this form does not correspond to wild-type yeast cells as the mutant expresses several filament-specific genes [285]. As yeast-hyphal transition is considered an invasive trait of *C. albicans*, gene expression induced by either the wild-type (hyphae) or *HGC1* null (yeast) strains was profiled and compared to unstimulated macrophages, as described previously [216]. We identified genes that showed significant differential expression in at least one of the conditions after Benjamini-Hochberg correction (P < 0.05 and > 2-fold change in expression). From this set, we selected the 62 genes that are exclusively induced after stimulation with wild-type *C. albicans* for 24h.

Candidemia and control cohorts

In this study, we included 227 unrelated adult Caucasian candidemia patients (described in detail in [280]). Patient enrollment took place after confirmation of at least one positive blood culture for a *Candida* species. The control cohort of 176 Caucasians consists of non-infected (candidiasis-free)
matched patients from the same medical centres as the patient cohort. Controls were recruited consecutively from the same hospital wards as infected patients during the study period, with a similar balance of medical, surgical, and oncology patients in case and control groups. Review boards of the involved medical centres approved the study and patients were enrolled after giving written informed consent (Supplementary Methods).

Genotyping and genetic analyses
Cases and controls in the candidemia cohort were genotyped on the Illumina Immunochip SNP array platform, which contains ~200,000 SNPs focused on genomic regions known to be involved in immune-mediated diseases [286]. After the application of quality control filters, we tested 64 SNPs in the FAP-IFIH1-GCA-KCNH7 LD region for candidemia case-control association. This 405 kb LD region (hg18 coordinates: chr2 162,720 kb – 163125 kb) was defined based on the LD patterns in the larger genomic region (Figure S1). See Supplementary Methods for detailed information.

Expression analysis of PBMCs
To assess the expression levels of genes in the FAP-IFIH1-GCA-KCNH7 LD region, blood was collected from healthy volunteers. Peripheral blood mononuclear cells (PBMCs, 5×10^5/well) were stimulated with either culture medium, 1 × 10^6/ml heat-killed Borrelia burgdorferi [287], 1×10^6/ml heat-killed Candida albicans (UC820) [288], 10 ng/ml Escherichia coli-derived lipopolysaccharide (LPS), or 1×10^7/ml sonicated Mycobacterium tuberculosis (MTB) (Hv37Rv) for either 4 or 24h. Gene expression was profiled using the Illumina Human HT-12 expression BeadChip [216]. Additionally, gene expression was assessed in PBMCs from two patients suffering from CMC due to STAT1 mutations (Arg274Trp) [282]. Cells were stimulated with C. albicans for 4h and gene expression was measured by RNA sequencing as described elsewhere [216].

PBMC stimulations experiments
5×10^5 isolated PBMCs per well were stimulated with either heat-killed C. albicans yeast or hyphae (UC820, 1×10^6/ml) for 24h (IL-10) or 7 days (IL-17 and IFN-γ). Additionally, PBMCs were stimulated with MTB (1µg/µl) for 24h. Supernatants were collected and measured for IL-10 and IFN-γ (Sanquin, Amsterdam, The Netherlands), and IL-17 cytokines (R&D Systems, Abingdon, UK). We excluded samples that showed a positive RPMI (control condition with contamination) or errors in the experimental readout.

Mda5 knockout mice studies
Mda5−/− mice on C57BL/6J background have been backcrossed at least ten times and were kindly provided by Dr. M. Colonna. Splenocytes were isolated from wild-type and Mda5−/− mice and stimulated with RPMI, Poly I:C, and heat-killed C. albicans yeasts or hyphae for 24h. Cytokines were measured in supernatants by ELISA. For quantitative real-time polymerase chain reaction (qPCR) experiments, splenocytes from both Mda5 knockout mice and B6 control mice were stimulated for 24h with heat-killed C. albicans hyphae. RNA was isolated according to the TRIzol® isolation protocol (Life Technologies).
Figure 1. Transcriptional changes in macrophages stimulated with Candida albicans. (A) The heatmap shows differential gene expression after 4h or 24h stimulation of human macrophages with yeast-locked HGC1 null C. albicans (which are unable to form hyphae, but are known to express several hyphal proteins), or wild-type invasive C. albicans (that can form hyphae), compared to expression levels in unstimulated macrophages (control). 62 genes exhibited a significant change in expression level (Benjamini-Hochberg-corrected P < 0.05 and > 2-fold change in expression) specifically after 24h stimulation with wild-type Candida, during which germination into hyphae takes place. The signal-to-noise ratio, scaled to the maximum absolute deviation, is shown for each probe corresponding to the 62 differentially expressed genes. (B) C. albicans hyphae-induced genes, IFIH1, TRIM25, ISG15 and IL8 (indicated in red), are components of the RIG-I-like receptor (RLR) signaling pathway. These genes represent both the MDA5 (IFIH1) and RIG-I (ISG15 and TRIM25) branches, as well as inflammatory cytokines that are produced by activation of the pathway (IL8). Figure based on the KEGG map of the RLR pathway [253].
**RESULTS**

*Candida* germination induces expression of RLR pathway components in macrophages

*Candida albicans* is a dimorphic fungus that exists either in a colonizing yeast form or as an invasive filamentous form (hyphae). To identify the specific transcription profile induced by fungal germination into hyphae, we profiled the transcriptome of macrophages stimulated with either wild-type (which develop hyphae) and *HGC1* null strains of *C. albicans* (which have a yeast-locked phenotype, though are different from wild-type) [285]. 62 genes exhibited significant differential expression specifically in macrophages stimulated with *Candida* hyphae for 24h (required for hyphal formation), but not for 4h (Benjamini-Hochberg-corrected *P* < 0.05 and > 2-fold change in expression compared to unstimulated macrophages, Figure 1A). Many of these genes are involved in interferon (IFN) signaling, consistent with a previous study [216]. Interestingly, four of the genes induced by *Candida* hyphae stimulation (IFIH1, ISG15, IL8, and TRIM25) are components of the RIG-I-like receptor (RLR) signaling pathway, significantly more than expected for a random set of genes (*P* = 4.3 × 10⁻³, 11.5-fold enrichment, Table 1).

RIG-I-like receptors are well-known intracellular receptors of viral RNA, leading to the production of type I IFNs and proinflammatory cytokines [2]. IFIH1, with its protein product known as MDAS, is the receptor of one branch of the RLR pathway (Figure 1B). ISG15 and TRIM25 are involved in the RIG-I branch. Thus, the invasive form of *Candida* induces expression of components of two branches of the virus-recognition RLR pathway in macrophages.

**Genetic variation linked to IFIH1 modulates susceptibility to candidemia**

To validate a role for components of the RLR pathway in invasive *Candida* infection, we investigated whether genetic variation linked to *IFIH1*, *ISG15*, *IL8*, or *TRIM25* correlates with susceptibility to candidemia in patients. Analysis of 64 SNPs associated with *IFIH1* revealed strong associations (Figure 2A and Table 2). The *IFIH1* locus is present in a 405 kb region on chromosome 2 with low recombination rates (Figure 2A) and accompanying strong linkage disequilibrium (LD) in both the candidemia cohort and the HapMap CEU population (Figures 2B and S1) [289-291]. Besides *IFIH1*, the

<table>
<thead>
<tr>
<th>KEGG identifier</th>
<th>Pathway</th>
<th><em>P</em> value</th>
<th>Fold enrichment</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa04622</td>
<td>RIG-I-like receptor signaling pathway</td>
<td>4.3 × 10⁻³</td>
<td>11.5</td>
<td>IFIH1, ISG15, IL8, TRIM25</td>
</tr>
<tr>
<td>hsa04060</td>
<td>Cytokine–cytokine receptor interaction</td>
<td>6.6 × 10⁻¹</td>
<td>4.7</td>
<td>OSM, TNFSF13B, IL8, CXCR4,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CCL3L1, CCL5</td>
</tr>
<tr>
<td>hsa04612</td>
<td>Antigen processing and presentation</td>
<td>6.6 × 10⁻¹</td>
<td>9.8</td>
<td>TAP2, TAP1, HSPA6, HSPA1B</td>
</tr>
<tr>
<td>hsa04062</td>
<td>Chemokine signaling pathway</td>
<td>1.1 × 10⁻²</td>
<td>5.4</td>
<td>IL8, CXCR4, CCL3L1, CCL5, STAT1</td>
</tr>
<tr>
<td>hsa04621</td>
<td>NOD-like receptor signaling pathway</td>
<td>3.4 × 10⁻²</td>
<td>9.8</td>
<td>IL8, RIPK2, CCL5</td>
</tr>
<tr>
<td>hsa05120</td>
<td>Epithelial cell signaling in Helicobacter pylori infection</td>
<td>4.0 × 10⁻²</td>
<td>9.0</td>
<td>IL8, HBEGF, CCL5</td>
</tr>
</tbody>
</table>

Enrichment for KEGG pathway components [253] was determined using the functional annotation tool of the DAVID suite [272]. Background: all human genes.
Table 2

Identifiers indicate the top SNP and two high D’ values. Haplotype blocks (triangles with bold black borders) are regions where at least 95% of SNPs are in strong LD, defined by which is calculated as D’ and LOD: SNPs with D’ values between 0 and 1 and with LOD ≥ 2 are colored from white to red. Immunochip SNP array. The intersections of the diagonals between pairs of SNPs are colored according to the degree of LD, from strong to weak LD - red: r² ≥ 0.8; orange: 0.5 ≤ r² < 0.8; yellow: 0.2 ≤ r² < 0.5; white: r² < 0.2. Genes with their direction of transcription are shown at the bottom; KCHN7 transcription are shown at the bottom;

Figure 2. (A) Regional association plot and (B) linkage disequilibrium (LD) map for the FAP-IFIH1-GCA-KCNH7 LD region on chromosome 2. (A) 64 SNPs with MAF > 5% in 403 Caucasian individuals of the candidemia cohort (cases and controls together) were assessed for genotypic association with candidemia. The resulting -log₁₀ (genotypic P values) (left y-axis) are plotted as a function of genomic coordinates (hg18, x-axis). The blue diamond highlights the most significant SNP along with its P value (rs984971). rs1990760 and rs3747517 are the only two significant missense SNPs; both are in the coding region of IFIH1. Recombination rates, estimated from the CEU, YRI, and JPT+CHB HapMap populations (HapMap 2, Release 22) [292], are plotted to reflect the local LD structure (right y-axis, cyan line). SNPs are colored according to the degree of LD with the most significant SNP, rs984971 (R-squared, calculated across the controls in the candidemia cohort; from strong to weak LD - red: r² ≥ 0.8; orange: 0.5 ≤ r² < 0.8; yellow: 0.2 ≤ r² < 0.5; white: r² < 0.2). Genes with their direction of transcription are shown at the bottom; KCNH7 is only partly in this region. (B) LD patterns across the 405 kb FAP-IFIH1-GCA-KCNH7 LD region are calculated based on genotypes of control individuals in the candidemia cohort, measured using the Immunochip SNP array. The intersections of the diagonals between pairs of SNPs are colored according to the degree of LD, which is calculated as D’ and LOD: SNPs with D’ values between 0 and 1 and with LOD ≥ 2 are colored from white to red. Haplotype blocks (triangles with bold black borders) are regions where at least 95% of SNPs are in strong LD, defined by high D’ values [281]. Chromosome 2 coordinates (hg18) and Entrez genes are shown at the top. Orange boxes around SNP identifiers indicate the top SNP and two IFIH1 missense SNPs significantly associated with susceptibility to candidemia (see Table 2). The corresponding R-squared LD map for the candidemia cohort is depicted in Figure S1D. See Figures S1A-C for R-squared and D’/LOD LD maps calculated based on the HapMap CEU population.
LD region contains the genes FAP, GCA, and part of KCNH7. 15 of the 64 SNPs in the FAP-IFIH1-GCA-KCNH7 LD region differ significantly between cases and controls (Benjamini-Hochberg-corrected genotypic $P < 0.05$, Table S1). The significant SNPs are distributed mainly across the central part of the LD region (Figure 2A) and the association does not extend beyond the LD region (Figure S2).

An intergenic SNP between GCA and KCNH7 shows the strongest association with candidemia (rs984971, genotypic $P = 2.2 \times 10^{-5}$, allelic $P = 2.2 \times 10^{-4}$, odds of disease 0.43 – 0.77, Table 2). Although the Immunochip covers missense coding variants in all four genes, only IFIH1 harbors...
significant missense SNPs (rs1990760 – Ala946Thr, and rs3747517 – His843Arg, which are also in strong LD with each other; HapMap CEU: D’ = 1, r2 = 0.42 – candidemia cohort: D’ = 1, r2 = 0.55). Furthermore, rs1990760 and rs3747517 are quantitative trait loci (QTLs) to IFIH1 expression, with the candidemia risk alleles correlating with higher IFIH1 expression in PBMCs [290,294]. These results suggest that IFIH1 is involved in candidemia.

**IFIH1 is strongly upregulated upon Candida stimulation of PBMCs, while FAP, GCA, and KCNH7 are not**

To provide additional evidence regarding which genes in the FAP-IFIH1-GCA-KCNH7 LD region are important for the host response to Candida, we assessed gene expression in PBMCs from healthy volunteers after stimulation with various microbes. Stimulation with *C. albicans* resulted in a strong increase of IFIH1 expression ($P = 1.5 \times 10^{-15}$ at 4h and $P = 1.9 \times 10^{-12}$ at 24h, Welch-corrected t-tests, Figure 3A). Of the other LD region genes, only GCA was also weakly induced by *C. albicans*. In addition, we compared the expression patterns of IFIH1 in healthy individuals with two patients suffering from chronic mucocutaneous candidiasis (CMC) due to a deleterious STAT1 mutation [282]. CMC patient cells expressed significantly lower levels of IFIH1 after stimulation with *C. albicans* than cells from healthy individuals ($P = 0.04$, Welch-corrected t-test, Figure 3B), while of the other genes, only GCA also displayed minor differences ($P = 0.05$). Together, the observations in healthy individuals and CMC patients indicate that expression of IFIH1, and not expression of the other genes in the LD region, is specifically induced by stimulation with *C. albicans*.

**Figure 3. Transcriptional response of genes in the FAP-IFIH1-GCA-KCNH7 LD region to various microbial stimuli.** (A) Peripheral blood mononuclear cells (PBMCs) from healthy volunteers (minimum n=23) were stimulated for either 4 or 24h with *Borrelia burgdorferi*, *Candida albicans*, *Escherichia coli*-derived lipopolysaccharide (LPS), or *Mycobacterium tuberculosis* (MTB). Gene expression (Mean ± SD) was measured using microarrays and normalized to the control RPMI condition (untreated). $P$ values (Welch-corrected t-test) compared expression distributions of individual stimuli to their respective untreated controls and were Bonferroni-corrected for testing 32 hypotheses (four stimuli across four genes at two time points). Asterisks represent all significant comparisons at $\alpha < 0.05$. (B) Gene expression (Mean ± SD) in PBMCs of healthy controls (n=3) and patients suffering from chronic mucocutaneous candidiasis (CMC) (n=2) were stimulated with *C. albicans* for 4 hours. $P$ values were calculated using the Welch-corrected t-test.
Genetic variants in *IFIH1* are associated with an altered cytokine profile in response to *Candida*

To investigate the functional consequences of genetic variants associated with *IFIH1* that predispose individuals to candidemia (Table 2), we correlated the genotypes of the SNPs with *in vitro* cytokines levels upon *Candida* stimulation. A trend was observed towards an increased capacity to release the proinflammatory cytokines IFN-γ and IL-17 in cells isolated from individuals homozygous for the risk allele for both *IFIH1* missense polymorphisms (TT for rs1990760; CC for rs3747517) (Figure 4). In contrast, levels of the anti-inflammatory IL-10 tended to be lower in individuals carrying the risk allele. The top intergenic SNP associated with candidemia (rs984971) did not reveal the same trends (Figure S3). Furthermore, stimulation with other microbial stimuli did not reveal clear correlations between cytokine levels and *IFIH1* missense SNP genotypes (Figure S4), suggesting specificity for *Candida*. Thus, genetic variation in *IFIH1* may influence anti-*Candida* cytokine profiles *in vitro*.

**Missense SNPs could affect MDA5 protein function**

We next sought to gain insight into the possible consequences of having alternative alleles at the *IFIH1* missense SNPs on MDA5 protein function. *In silico* analysis shows that residue 946 (rs1990760, Ala946Thr) is part of an intrinsically disordered loop [239]. The equivalent loop is rigid in RIG-I, and this differential flexibility contributes to the different RNA binding preferences between MDA5 and RIG-I [295]. The human MDA5 crystal structure has an arginine at position 843

![Figure 4. IFIH1 missense SNP genotypes correlate with Candida-induced cytokine levels.](image)
(rs3747517, His843Arg), which interacts with the negatively charged RNA backbone (Figure S5). Histidine would weaken this electrostatic interaction because it is less often positively charged at physiological pH than arginine. Furthermore, position 843 is close to the interface likely involved in interactions between MDA5 monomers (Figure S5) [295]. The formation of MD5 filaments along the RNA is critical for downstream activation of MAVS [296] and mutation of nearby residues 841 and 842 disrupts signaling [295]. Thus, the Ala946Thr and His843Arg substitutions could alter dsRNA binding selectivity and affinity, and the latter might also affect signaling activity.

**Mda5 knockout mice have reduced cytokine production in response to C. albicans**

To provide an additional argument for the role of MDA5 in the anti-Candida response, we stimulated splenocytes from Mda5 knockout and B6 control mice with C. albicans yeasts or hyphae. Mda5-deficient cells showed a defective production of interferon β induced by *Candida* (Figure 5). Similarly, the IL-6 and IL-10 cytokine responses were lower in cells from Mda5 knockout mice. These differences were more pronounced in stimulations with hyphae compared to the yeast form (Figure S6).

**Discussion**

In the present study, we propose that the pattern recognition receptor MD5, which belongs to the RIG-I-like receptor (RLR) family and plays an important role in antiviral immunity by recognizing viral RNA [2], is also involved in antifungal host defense. MD5 modulates cytokine production induced in human leukocytes by *C. albicans*, while genetic variants in the IFIH1 gene that encodes MD5 influence susceptibility to disseminated candidiasis. Based on these data and the known role of MD5 in interferon (IFN) production, it is most likely that this effect is mediated through the induction of type I IFNs.

*C. albicans* is a dimorphic fungus, and germination from yeasts to hyphae is a central process for the invasion of tissues. Surprisingly, transcriptome analysis aiming to identify the immunological programs induced in human macrophages specifically by *Candida* germination into hyphae identified the MD5/RIG-I signaling pathway as one of the top targets. The hypothesis that MD5 is important for host defense against *Candida* was strengthened by the observation that MD5 induction is defective in cells isolated from patients suffering from chronic mucocutaneous candidiasis. Furthermore, analysis of the genomic region that contains IFIH1 in a cohort of candidemia patients revealed a strong, though not genome-wide significant, association between genetic variation occurring in this genomic region and the disease. While genetic variation in
IFIH1 has previously been shown to influence susceptibility to several autoimmune diseases such as type I diabetes, Graves’ disease, and multiple sclerosis [290,291,294,297-302], this is the first report of polymorphisms in IFIH1 linked to a fungal infection. These data are in line with recent studies showing that polymorphisms in other pattern recognition receptors such as TLRs [280,303-305], or components of the IFN pathway such as STAT1 or IRF1 [216], also influence susceptibility to systemic fungal infections.

It is important to point out that the candidemia-associated LD region contains several genes: FAP (fibroblast activation protein), IFIH1 (interferon induced with helicase C domain 1), GCA (grancalcin), and KCNH7 (potassium voltage-gated channel subfamily H member 7). IFIH1 and grancalcin were the strongest candidates for causing the susceptibility to candidemia, as these genes have known functions in immunity. Grancalcin is abundant in macrophages and neutrophils [306], and is thought to mediate leukocyte adhesion and migration [307]. Gene expression analysis confirmed that IFIH1 was strongly induced in PBMCs stimulated with Candida and GCA to lesser extent, while the other genes did not show any expression changes. IFIH1 and GCA are divergently transcribed neighboring genes with ~25kb separating their transcription start sites. As such neighboring genes tend to be co-expressed [308,309], the moderate upregulation of GCA in response to Candida stimulation could be a by-effect of the strong induction of IFIH1, although the genes are still relatively far apart. Importantly, grancalcin-deficient (Gca−/−) mice are not more susceptible to candidiasis than wild-type mice [310], which strongly argues against an important role for GCA in the immune response against C. albicans. Therefore, we concluded that genetic variants acting at IFIH1 are the most likely cause of the association of the FAP-IFIH1-GCA-KCNH7 LD region with candidemia.

The candidemia risk alleles of the IFIH1-linked SNPs identified in our study have previously been shown to lead to higher expression of IFIH1 in PBMCs [290,294]. Furthermore, our protein structure analysis indicates a possibly stronger RNA binding by MDA5 through Arg843, which is encoded by the risk allele C of rs3747517. To gain insight into the downstream immunological effects of IFIH1 variants, we measured cytokine levels produced by PBMCs with different genotypes for the two IFIH1 missense SNPs (rs1990760 and rs3747517). These data indicate that cells from individuals bearing the candidemia risk alleles produce more proinflammatory cytokines (IFN-γ and IL-17) and less anti-inflammatory IL-10 in response to C. albicans yeast and hyphal forms than cells bearing the protective alleles.

These observations bring into discussion the nature of the involvement of MDA5 in the host defense against Candida. MDA5 activates the RLR pathway, leading to the production of type I IFNs during viral infections [202]. A similar biological activity during Candida stimulation was shown by our data from Mda5 knockout mouse splenocytes, which displayed a decreased capacity to induce IFN-β. A role for type I IFNs in antifungal immunity has been recently proposed [216], and MDA5 is likely the receptor that is at least partially responsible for the type I IFN induction during C. albicans infection.

Mutations leading to inherently increased expression or activity of MDA5 are likely to increase IFN production [311]. Aberrant production of type I IFNs in turn can cause imbalances in the immune response that are reflected in our observed alterations in the levels of other cytokines. The
apparent deleterious effect of MDA5 hyperactivity on the anti-
Candida host defense is consistent with observations that type I IFNs could be harmful for this response: mice defective in type I IFN receptors (Ifnar1−/− mice) are actually more resistant to systemic Candida infections [312]. This is also in line with our findings that PBMCs with the candidemia risk genotype in IFIH1 tend to release more inflammatory cytokines. The hypothesis that MDA5 has a negative effect on the anti-
Candida immune response has been proven by a very recent elegant study demonstrating that Mda5−/− mice are more resistant to disseminated candidiasis (Malireddi and Kanneganti, personal communication).

It is currently unclear which ligands cause activation of MDAS in Candida infection. Candida is mainly recognized by cell surface pattern recognition receptors such as TLRs and C-type lectin receptors (CLRs), after which the fungus is internalized and subsequently digested in the phagolysosome [313,314]. It is conceivable that, during this process, Candida-derived structures may leak from these organelles and enter the cytoplasm, a process described earlier for the recognition of mycobacterial peptidoglycans by the cytoplasmic receptor NOD2 [315-317]. Interestingly, a recent study has suggested that NOD2 is also important for the recognition of Candida chitin [318]. Nevertheless, there is currently no experimental evidence to support that either the wild-type form or a variant form of MDA5 has ligands other than the described RNAs.

In conclusion, this study demonstrates that the viral receptor MDA5 has an important role in modulating innate immune responses against the fungal pathogen C. albicans. Future research should shed light on the exact mechanisms through which MDA5 participates in the defense against the fungus. Nevertheless, the possible deleterious effects of MDA5-dependent stimulation during systemic candidiasis shown by our data suggest its potential usefulness as a novel therapeutic target.

ACKNOWLEDGMENTS

We thank Martin Oti for helpful discussions. MJ and MGN were supported by an ERC Consolidator Grant (nr. 310372 to MGN). RvdL and MAH were supported by the Virgo consortium, funded by the Dutch government (FES0908), and by the Netherlands Genomics Initiative (050-060-452). CW was supported by the ERC Advanced Grant, ERC-671274. XW was supported by NSFC 11101321 and 61263039 grants.
Supplementary Methods

Genotyping and quality control (extended)
DNA was isolated from whole blood using the Gentra Pure Gene Blood kit (Qiagen, Venlo, The Netherlands), according to the protocol of the manufacturer. Only samples with a SNP call rate above 90% were included. We applied quality control filters to exclude SNPs with: (i) a genotype call rate of less than 90%, (ii) strong deviation from Hardy-Weinberg equilibrium in control samples (Hardy-Weinberg exact test $P \leq 1 \times 10^{-3}$), and (iii) significant differences in missingness between cases and controls (Fisher’s exact $P < 1 \times 10^{-2}$). As our cohort does not have the power to detect associations with rare variants, we only included SNPs with a minor allele frequency (MAF) of greater than 5% in cases and controls together.

Genetic analyses (extended)
Linkage disequilibrium (LD) measures were calculated using Haploview version 4.2 [319] for both the candidemia cohort (LD patterns are based on genotypes of control individuals only) and common SNPs (MAF > 5%) in the CEU population (Utah Residents (CEPH) with Northern and Western European ancestry) in HapMap 3, release 2 [158]. The 405 kb FAP-IFIH1-GCA-KCNH7 LD region on chromosome 2 (hg18 coordinates: 162,720 kb – 163125 kb) was defined based on the LD patterns in the larger genomic region (Figure S1). It includes the complete FAP gene and part of the KCNH7 gene.

Associations between SNPs and susceptibility to candidemia were assessed using both genotypic and allelic tests. Genotypic association was calculated using the Fisher’s exact test, asking whether candidemia cases and controls have significantly different genotype count distributions ($H_0$: genotype counts are the same). We corrected for testing multiple SNPs using the Benjamini-Hochberg procedure. Similarly, allelic association was calculated using the Fisher’s exact test, asking whether candidemia cases and controls have significantly different allele count distributions ($H_0$: allele counts are the same). Odds ratios (OR, with 95% confidence intervals) are reported for the allelic association tests and represent the odds of disease for individuals carrying the non-risk allele versus the risk allele. Quality filtering and genetic analyses were performed using PLINK v1.07 [320] and custom R scripts. Regional association plots were created using code adapted from http://www.broadinstitute.org/diabetes/scandinavs/figures.html.

Ethics statement
PBMCs were isolated from blood of healthy volunteers after obtaining written informed consent. Candidemia patients were enrolled after giving written informed consent at the Duke University Hospital (Durham, North Carolina, USA) as well as the Radboud University Medical Centre (Nijmegen, NL). The study was approved by the Institutional Review Boards of both medical centres, the Institutional Review Board of Duke University (CR4_Pro00006427) and the ‘Commissie Mensgebonden Onderzoek Arnhem-Nijmegen’ (2001/198). The study was performed in accordance with the international guidelines of the declaration of Helsinki (year 2000) of the World Medical Association adopted by the World Medical Assembly. Enrollment took place between January 2003 and January 2009. Mice studies were reviewed and approved by the internal animal care committee of the University Hospital Bonn and were performed according to national and European regulations.
Supplementary Figures
Figure S1 (previous page). Linkage disequilibrium (LD) maps of the FAP-IFIH1-GCA-KCNH7 LD region on chromosome 2. (A) 1.5 Mb region centered on the FAP-IFIH1-GCA-KCNH7 LD region. (B) 525 kb region centered on the FAP-IFIH1-GCA-KCNH7 LD region (which is shaded in green in panel A). (C-D) The 405 kb FAP-IFIH1-GCA-KCNH7 LD region itself (which is shaded in purple in panel A, and marked by purple dashed lines in panel B). (A-C) LD patterns are calculated based on 293 (panel A), 59 (B), and 31 (C) common SNPs (minor allele frequency, MAF > 5%) genotyped in 117 individuals from the CEU population (Utah Residents (CEPH) with Northern and Western European ancestry) in HapMap 3, release 2 [158]. (D) LD patterns are calculated based on genotypes of control individuals in the candidemia cohort for 64 SNPs with MAF > 5% (cases and controls together), measured using the Immunochip SNP array. The intersections of the diagonals between pairs of SNPs (if separated by less than 500 kb) are colored according to the degree of LD. The middle part of each panel shows R-squared (r2; correlation coefficient between genotypes): r2 = 0 (no correlation) is in white, r2 = 1 (complete correlation) is in black, and r2 values between 0 and 1 are in increasing intensities of gray. The bottom part of each panel shows D' and LOD: SNPs with D' values between 0 and 1 and with LOD ≥ 2 are colored from white to red (D' = 1 suggest that no recombination has occurred between two SNPs, thus indicating complete LD; logarithm of odds (LOD) scores ≥ 2 suggest that two markers are inherited together). Haplotype blocks (triangles with bold black borders) are regions where at least 95% of SNPs are in strong LD, defined by high D' values [293]. Chromosome 2 coordinates (hg18) and Entrez genes in these genomic regions are shown in the top part of each panel. Orange boxes around SNP identifiers indicate the top SNP and two IFIH1 missense SNPs significantly associated with susceptibility to candidemia (see Table 2 in the main text). Figures created using Haploview version 4.2 [319].

Figure S2. Regional association plot for the 10 Mb region around the FAP-IFIH1-GCA-KCNH7 LD region on chromosome 2. 203 SNPs in this region with MAF > 5% in 403 Caucasian individuals of the candidemia cohort (cases and controls together) were assessed for genotypic association with candidemia. The resulting -log_{10}(genotypic P values) (left y-axis) are plotted as a function of genomic coordinates (hg18, x-axis). The blue diamond highlights the most significant SNP with its P value (rs984971). Recombination rates, estimated from the CEU, YRI and JPT+CHB HapMap populations (HapMap 2, Release 22 [282]), are plotted to reflect the local LD structure around the associated SNPs (right y-axis, cyan line). SNPs are colored according to the degree of LD with the most significant SNP, rs984971 (R-squared, calculated across the controls in the candidemia cohort; from strong to weak LD - red: r²≥0.8; orange: 0.5≤r²<0.8; yellow: 0.2≤r²<0.5; white: r²<0.2). Genes with their direction of transcription are shown at the bottom. The association with candidemia in this genomic region is limited to the FAP-IFIH1-GCA-KCNH7 LD region (marked by purple dashed lines), although not all regions are as densely genotyped by the Immunochip array.
Figure S3 (left). Genotypes of the intergenic SNP rs984971 do not show the same Candida-induced cytokine level trends as the missense SNPs rs1990760 and rs3747517. PBMCs from healthy volunteers with different genotypes for rs984971 (candidemia risk allele A) were stimulated in vitro with either C. albicans yeast or hyphae. Cytokine levels (scatterplots with mean indicated) were measured after 24 hours (IL-10) or 7 days (IL-17 and IFN-γ) by enzyme linked immunosorbent assay (ELISA). P values were calculated using the Mann-Whitney U test comparing cytokine levels of the two homozygous genotypes.

Figure S4 (right). IFIH1 missense SNP genotypes do not correlate with MTB-induced cytokine levels. PBMCs from healthy volunteers with different genotypes for rs1990760 and rs3747517 were stimulated in vitro with Mycobacterium tuberculosis. Cytokine levels (scatterplots with mean indicated) were measured after 24 hours (IL-10) or 7 days (IL-17 and IFN-γ) by enzyme linked immunosorbent assay (ELISA). P values were calculated using the Mann-Whitney U test comparing cytokine levels of the two homozygous genotypes.

Figure S6 (left). Mean (+SEM) in vitro cytokine response of mouse splenocytes after stimulation with different stimuli. Splenocytes from B6 control mice and Mda5 knockout mice were stimulated for 24 hours. (A) mIL-6 levels (B) mIL-10 levels.
Figure S5. rs3747517 (His843Arg) could affect RNA binding and MDA5 filament formation. Arginine 843 (R843, red) is depicted in the crystal structure of MDA5 Δ2CARD (gray) bound to dsRNA (blue) (PDB 4GL2) [295]. R843 is adjacent to the negatively charged RNA backbone and close to the likely protein-protein interaction interface (green dashed line) involved in MDA5 oligomerization. Molecular graphics created with YASARA (www.yasara.org).

Residue 946 (rs1990760, Ala946Thr) is part of an intrinsically disordered loop [239] in the C-terminal domain (CTD) of MDA5. The equivalent loop is rigid in RIG-I, and this differential flexibility contributes to the different RNA binding preferences: displacement of the loop upon dsRNA binding allows MDA5 to recognize long dsRNA, while the loop in RIG-I specifically caps shorter dsRNA [295].

MDA5 arose from a duplication of the MDA5/LGP2 gene in the ancestor of jawed vertebrates [321]. Analysis of MDA5 orthologs in 59 jawed vertebrates reveals that both amino acids of rs1990760 (Ala946Thr) are common (alanine: 36/59=61%, threonine: 10/59=17%), while for rs3747517 (His843Arg) the type sequence of human is the only sequence that encodes a histidine (similar data in [322]). The occurrence of alternative alleles at both SNPs suggests that both lead to a functional protein, although there may be functional differences.
Table S1. SNPs in the FAP-IFIH1-GCA-KCNH7 LD region are significantly associated with susceptibility to candidemia.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Immunochip</th>
<th>Chromosome 1 coordinate</th>
<th>Allele 1 (dbSNP)</th>
<th>Allele 2 (dbSNP)</th>
<th>Allele frequencies</th>
<th>Allelic OR (95% c.i.)</th>
<th>Allelic P-value</th>
<th>Genotypic OR (95% c.i.)</th>
<th>Genotypic P-value</th>
<th>Genotypic value</th>
<th>BH-corrected genotypic P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs984971</td>
<td>imm_2_162932767</td>
<td>162,932,767</td>
<td>GCA intergenic</td>
<td>163,224,521</td>
<td>G</td>
<td>A</td>
<td>153/199</td>
<td>2.2E-04</td>
<td>4/57/117</td>
<td>25/89/113</td>
<td>2.2E-05</td>
</tr>
<tr>
<td>rs2068330</td>
<td>imm_2_162945636</td>
<td>162,945,636</td>
<td>KCNH7 intronic</td>
<td>163,237,390</td>
<td>G</td>
<td>C</td>
<td>153/199</td>
<td>2.2E-04</td>
<td>4/57/117</td>
<td>25/89/113</td>
<td>2.2E-05</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

SNPs in the FAP-IFIH1-GCA-KCNH7 LD region with MAF > 5% in cases and control together were assessed for genotypic and allelic association with candidemia using the Fisher’s exact test. Immunochip: the identifier of the SNP on the Immunochip SNP array. Alleles: the alleles measured with the Immunochip, and the complementary alleles reported by dbSNP (build 138), if they are different. Odds ratios (OR, with 95% confidence intervals) are reported for the allelic association tests and represent the odds of disease for individuals carrying Allele 1 versus Allele 2. Allele frequencies columns report the occurrence of Allele 1, followed by Allele 2. Genotype frequencies columns report the number of homozygotes for Allele 1, followed by the number of heterozygotes, followed by the number of homozygotes for Allele 2. BH-corrected genotypic P-value: Benjamini-Hochberg correction of the genotypic association test P values for testing multiple SNPs (64 in total). The table shows all SNPs significantly associated with candidemia (BH-corrected genotypic P < 0.05). These SNPs are in strong LD in both the HapMap CEU population (average pairwise D' = 0.94, average pairwise r2 = 0.43, Figure S1C) and in the candidemia cohort (average pairwise D' = 0.91, average pairwise r2 = 0.41, Figure S1D). All SNPs tested are in Hardy–Weinberg equilibrium in the controls (P > 1 × 10^-3).
Chapter 4

Disease gene prioritization in recurrent herpes virus infections using systems immunology

This chapter summarizes our contribution to, and is partly adapted from:

Immunologic defects in severe mucocutaneous HSV-2 infections: Response to IFN-γ therapy

Journal of Allergy and Clinical Immunology (2016), 138 (3): 895-898

Peer Arts¹, Frank L. van de Veerdonk², Robin van der Lee³, Martijn A. Langereis⁴, Christian Gilissen¹, Wendy A.G. van Zelst-Stams¹, Martijn A. Huynen³, Jos W.M. van der Meer², Frank J. van Kuppeveld⁴, Joris A. Veltman¹,⁵,⁶, Bart Jan Kullberg², Alexander Hoischen¹,⁵, Mihai G. Netea²

¹Department of Human Genetics, ²Department of Internal Medicine and Radboudumc Center for Infectious Diseases, ³Centre for Molecular and Biomolecular Informatics – Radboud university medical center, Nijmegen, The Netherlands ⁴Virology Division, Department of Infectious Diseases and Immunology, Faculty of Veterinary Medicine, Utrecht University, NL ⁵Donders Centre for Cognitive Neuroimaging, Radboud University Nijmegen, NL ⁶Department of Medical Genetics, Maastricht University Medical Centre, NL
Background
Herpes viruses such as herpes simplex virus type 1, type 2, and varicella zoster virus (HSV-1, HSV-2 and VZV) cause lesions of the oral mucosa and skin, but can also affect other organs [323]. While most herpes infections are self-limiting, reactivations of latent herpes viruses have been reported in patients with genetic defects of the immune system such as in primary immunodeficiencies [324]. A crucial protective role has been demonstrated for interferon signaling pathways and innate pattern recognition genes such as TLR3, TRIF and TRAF3 [324]. Despite these findings, the mechanisms behind increased susceptibility to herpes infections remain unknown in most patients.

Our collaborators investigated three patients with severe recurrent HSV-2 skin infections (Figure 1). Immunological assessment of white blood cells revealed defects in double-stranded RNA recognition (polyIC dsRNA; a virus-like TLR3 ligand) and interferon-γ (IFNγ) production as the underlying cause of the susceptibility. Standard antiviral treatment was ineffective, but in contrast all patients responded with impressive resolution of their symptoms upon replacement therapy with recombinant IFNγ.

Prioritizing disease variants in HSV-2 patients
To elucidate the genetic basis of the HSV-2 infection susceptibilities, our collaborators performed whole-exome sequencing of the three patients. Analyzed variants were restricted to those that (i) affect canonical splice sites or lead to nonsynonymous substitutions, and (ii) are very rare (population frequency <0.25%) in both worldwide dbSNP data and in >5,000 in-house exomes. This filtering resulted in 164 (patient 1), 167 (patient 2) and 280 (patient 3) variants. No variants were identified in the same gene in all three patients, suggesting that the underlying immunodeficiency may be genetically heterogeneous.

To prioritize the identified rare genetic variants for a possible role in the disease phenotype, we systematically investigated overlaps between patient genes (i.e. genes containing variants in the HSV-2 patients) and genes involved in the innate antiviral immune response. We took two approaches. First, we analyzed known pathways involved in herpes virus recognition such as the cytosolic RNA/DNA sensing and Toll-like receptor pathways [325]. Second, we examined our resource of predicted components of innate antiviral signaling, obtained through integration of various genomics data (Chapter 2 [326]). The prioritization steps revealed 23 (patient 1), 29 (patient 2) and 33 (patient 3) variants associated with immunological pathways (Table 1). For each patient, we then investigated the candidate variants that best matched the observed immune defects.

Investigating plausible disease-causing variants
Exome analysis of patient 1 and her healthy parents identified two compound heterozygous variants in Histone deacetylase 5 (HDAC5; maternally-inherited p.E836Q, paternally-inherited p.P886H; both predicted to be damaging; Figure 1A). HDAC5 was reported to regulate the inflammatory response in macrophages and influence anti-HSV-1 immunity [327]. Patient 1 also carries a missense variant in the NOD-like receptor X1 (NLRX1 p.R860Q). NLRX1 modulates antiviral RIG-I-MAVS, NF-κB and reactive oxygen species (ROS) signaling [328].
Herpes infections in multiple family members of patient 2 suggest a shared defect. Exome sequencing of an affected niece revealed 34 variants that overlap with patient 2. One of those is a very rare variant in Zinc finger and BTB domain-containing 25 (ZBTB25 p.S7G; minor allele frequency = 0.004% [329]). ZBTB25 has been reported as a T cell-enriched transcription factor (TF) that negatively regulates the Nuclear factor of activated T cells (NF-AT) [330], which in turn is an important regulator of IFNγ expression by T-cells [331]. Based on these studies it is likely that ZBTB25 plays a role in regulating IFNγ production, which we speculate may be the primary explanation for the IFNγ defect observed in this patient. A second potential mechanism by which ZBTB25 may contribute to infection susceptibility is through innate antiviral pathways producing type I IFNs.

### Table 1. Rare genetic variants in HSV-2 patients that are associated with immunological pathways

<table>
<thead>
<tr>
<th>Patient 1</th>
<th>Gene</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZEB1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VEGFA</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD27</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIRT1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NUCCD1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLRX1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRKCH</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAB5C</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HDAC5</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDC6A8</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patient 2</th>
<th>Gene</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>TLR1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IL5, SLC11A1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IFIH1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLRX5</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JAK1, PTPN6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLEC(2x)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLA2G6</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTC7A</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENPP3, OPRK1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AEN, IFI44</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EIF3E *</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LRBA</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FHL2 *</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZBTB25 *</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patient 3</th>
<th>Gene</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>AQP9</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IL27RA</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CIITA</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IL6ST</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZFPM1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UNC13D</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LGALS3BP</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BPI, PLEC1D</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IL32</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POLR3B</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DDX41</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANXA2, GAPDH</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POLR3E</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAMA4, PLA2G3, THBS4</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STXB2P2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NUP153, PLEC</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Patient genes were prioritized based on overlaps with immune response genes. Analyses A-E were mostly done by our collaborators; we contributed F and G. A = Mouse KO "immune system phenotype"; B = Gene Ontology "immune"; C = NCBI gene search "interferon"; D = KEGG class "infectious disease"; E = immunodeficiency diagnostics; F = predicted components of innate antiviral signaling based on e.g. human PPIs, pathogen-human PPIs, and innate immune TFs (Chapter 2 [326]); G = innate cytosolic RNA/DNA sensing pathways.

The most promising candidate variants (green) are discussed in the main text. Other notable variants that may contribute to the disease include: JAK1 (patient 2), a critical component of IFNα/β and IFNγ signaling through the JAK-STAT pathway; and IFNG (patient 2 niece). Reported variants have been validated by Sanger sequencing. * denotes variants in patient 2 also carried by the affected niece. See Table E4 for all variants and further details (Online Repository at the Journal of Allergy and Clinical Immunology website, http://dx.doi.org/10.1016/j.jaci.2016.02.025).
which are important in herpes virus infections [325]. Indeed, patients 1 and 2 also showed a 30-50% decrease in IFNα/β production. ZBTB25 is among the 94 novel genes suggested to play a role in antiviral RIG-I-like (RLR) receptor signaling by our work in Chapter 2 [326]. Features indicating an antiviral function for ZBTB25 included interactions with the RNA/DNA sensing pathway (TRAF2, MAPK9) and the presence of conserved motifs for antiviral TFs (AP-1 and NF-κB) in its promoter. Following these findings, our collaborators genotyped the ZBTB25 variant in other family members and found complete co-segregation with the susceptibility to herpes infections: all four affected individuals are heterozygous for p.S7G ZBTB25, while all three tested healthy individuals are homozygous for the wild-type allele (Figure 1B). An additional missense variant affecting IFIH1 (MDA5, a RIG-I-like receptor of viral dsRNA; p.R77W) possibly explains the more severe phenotype of patient 2 compared to affected relatives (Figure 1B), as mutations in IFIH1 have been reported to affect IFN production [311]. Thus, the ZBTB25 variant is a promising candidate for contributing to HSV-2 susceptibility, given its co-segregation with the clinical phenotype, its rarity in the general population, and the agreement with the molecular and immunological picture in the patients.

Figure 1. Pedigrees of the three patients with recurrent HSV-2 infections, showing segregation of candidate genetic variants. Arrows indicate the index patients (in black). (A) Patient 1 (woman, 39) suffered from recurrent HSV-2 skin infections since age 3 months, varying in frequency from monthly to weekly. Exome analysis was performed on the patient and her healthy parents. (B) Patient 2 (woman, 59) suffered from shingles at age 11 and recurrent HSV-2 infections every 3-4 weeks since age 41. Several relatives (in gray) also had reported herpes virus infections. Exome sequencing was performed on the patient (III.6) and an affected niece (IV.2). Co-segregation analysis of candidate variants was performed in six additional family members. (C) Patient 3 (woman, 49) suffered from recurrent HSV infections since age 16. No DNA was available for co-segregation analysis.
Patient 3 carries rare variants in two different subunits of the RNA polymerase III complex (POLR3B p.P162A, POLR3E p.R671Q; Figure 1C). This complex detects DNA viruses such as Herpesviridae by sensing viral dsDNA in the cytosol and transcribing it to dsRNA, which in turn may be recognized by the RLR pathway, both leading to type I IFN production [325,332].

Discussion
Exome sequencing typically identifies hundreds of candidate genetic variants associated with a disease, even after filtering for rare and protein-changing variants. On top of this, many diseases have no single genetic cause but are heterogeneous defects of a larger biological system [193]. Prioritization approaches are thus required to identify causal variants, with knowledge-based analysis by experts and data-driven computational pipelines [10,16,193] representing two extremes. While fully automated methods have wider applicability, intermediate approaches between data-driven and knowledge-based are likely to provide better predictions; i.e. they provide sets of genes predicted to play a role in the relevant system based on the latest genomics data. Here we applied our resource of innate antiviral response genes (Chapter 2), based on both genomics data and knowledge of the immunity system, to prioritize disease variants in patients with severe recurrent HSV-2 infections. Our predictions were principally aimed at cytosolic RNA sensing pathways, which have a large overlap with the DNA sensing pathways that initiate the main innate response to herpes viruses [325]. In conclusion, our approach has identified promising candidate genetic variants that potentially contribute to herpes infection susceptibilities. Additional studies (more patients, deeper molecular characterization) are warranted to establish causality.
Genome-scale detection of positive selection in 9 primates predicts human-virus evolutionary conflicts

Manuscript in preparation

Robin van der Lee
Teunis J. P. van Dam
Martijn A. Huynen

Centre for Molecular and Biomolecular Informatics,
Radboud Institute for Molecular Life Sciences,
Radboud university medical center, Nijmegen, The Netherlands
**Abstract**

Hotspots of rapid genome evolution hold clues about human-virus genetic conflicts. In this study we present a large-scale comparative analysis of nine whole-genome sequenced primates, spanning 36–50 million years of evolution, to identify conservative, high-confidence targets of positive selection involved in the virus-host interaction. We find statistical evidence for positive selection acting on 331 human protein-coding genes (3% of 11,096 ortholog clusters), pinpointing 934 adaptively evolving codons. Our stringent pipeline overcomes limitations that have substantially affected previous estimates of positive selection: rigorous quality control on the alignments and evolutionary inferences revealed artefacts in 20% of 416 genes with initial evidence for positive selection. The artefacts are caused by systematic issues in transcript definitions (61% of problem cases), inconsistent gene models (38%), orthology inference (9%), and misalignment (5%). We highlight examples where apparent adaptive evolution of codons was incorrectly reported due to alignment of pseudogenes, alternative 5’/3’ ends of exons and genes, or tandem duplicated exons, although these may be adaptive evolutionary innovations in their own right. Our final 331 positively selected genes (PSG) are strongly enriched for innate and adaptive immune functions, secreted and cell membrane proteins (e.g. pattern recognition, complement and cytokine pathways, defensins, immune cell receptors, MHC, Siglecs). We also find evidence for positive selection in genes related to reproduction, chromosome segregation and meiosis (e.g. centromere-associated proteins CENPO, CENPT), dietary diversity (smell and taste receptors), and cholesterol/lipid transport (apolipoproteins). Focusing on the virus-host interaction, our PSG retrieve most known cases of evolutionary conflicts that influence antiviral activity (e.g. TRIM5, MAVS, SAMHD1, tetherin). We then predict 70 novel virus-human genetic conflicts by integrating the positively selected genes with genomic data describing the virus-host interaction (virus-human PPIs, immune cell expression, infection RNAi screens). Finally, through protein structure analysis we identify positively selected positions directly involved in the interaction interfaces between viruses and their human entry receptors (CD4 – HIV; CD46 [MCP] – measles, adenoviruses; CD55 [DAF] – picornaviruses). Our high-quality dataset of positively selected genes and residues in primates is a rich source for studying the genetics underlying human (antiviral) phenotypes.

**Introduction**
Conservation of structure and sequence often indicate biological function. Rapidly evolving sequence features may however also hold insights and point to molecular adaptations to selection pressures. What drives rapid genetic changes during evolution? Can these changes explain the specific phenotypes of species or individuals, such as a differential susceptibility to viruses?

Immunity genes contain the strongest signatures of rapid evolution due to positive Darwinian selection [59,62-66,68,326,333,334]. Pathogens continuously invent new ways to evade, counteract and suppress the immune response of their hosts, thereby acting as major drivers of the observed adaptive evolution of immune systems [58,69]. Numerous proteins involved in the virus-host interaction have been demonstrated to be in genetic conflict with their interacting viral proteins,
Positive selection in the human-virus interaction

Positive selection can be detected through comparative analysis of protein-coding DNA sequences from multiple species [59,60]. Markov models of codon evolution combined with maximum likelihood (ML) methods [61] can analyze alignments of orthologous sequences to identify genes, codons and lineages that show an excess of nonsynonymous substitutions (mutations in the DNA that cause changes to the protein) compared to synonymous (‘silent’) substitutions ($d_\text{N}/d_\text{S}$ ratio or $\omega$, see Text S1 for a detailed explanation). Successful application requires many steps [69,335](Text S1), including: (i) the identification of orthologous sequences, sampled from species across an appropriate evolutionary distance (distant enough to show variation, but not too divergent to prevent saturation); (ii) accurate alignment and phylogenetic tree reconstruction; (iii) parameterization of the ML model. While studies of positive selection on single genes have achieved reliable results, estimates of positive selection in whole genomes have been substantially affected by unreliabilities in sequencing, gene models, annotation and misalignment [46,336-341].

In this study, we performed comparative evolutionary analyses of recent whole-genome sequenced primates to identify conservative, high-confidence targets of positive selection. Our findings provide insights into the biological systems that have undergone molecular adaptation in primate evolution. Integration of the positively selected genes with structural and genomic data describing virus-host interactions provides insights into the determinants of viral infection and predicts new virus-human evolutionary conflicts.

**Results**

To obtain a confident dataset of positively selected genes relevant to human biology, we investigated protein-coding DNA sequences from nine simian (‘higher’) primates for which whole-genome sequences are available (Table S1, five genomes released as recent as 2012, three of which in 2014 [342]). This set includes five hominoids (‘apes’; human, chimpanzee, gorilla, orangutan, gibbon), three Old World Monkeys (macaque, baboon, vervet) and one New World monkey (marmoset), spanning an estimated 36–50 million years of evolutionary divergence [343].

**A reliable pipeline for conservative inference of positive selection**

Given the high incidence of false positives in large-scale detection of positive selection reported in literature [46,336-341], we developed a stringent six-stage pipeline that we subjected to rigorous curation and quality control at all steps (Figure 1, Methods). To minimize artefacts, we prioritized
To limit the influence of evolutionary processes (e.g. gene duplications), we only assessed clusters of one-to-one orthologous codons. To reduce alignment of nonhomologous codons, we computed multiple sequence alignments using PRANK. To achieve maximum alignment quality, we masked low-confidence codons and columns, and filtered out entire alignments based on the GUIDANCE and TCS algorithms, two distinct concepts for assessing reliability. These steps resulted in 11,096 alignments, representing half of the human protein-coding genes. Detailed inspection revealed that the alignments are of good overall quality, with the major improvements coming from PRANK. The masking procedures filter out most of the remaining ambiguities.

Next, we used $d_N/d_S$-based codon substitution maximum likelihood (ML) models to infer positive selection acting on genes and codons. This requires an estimation of the overall evolutionary divergence between the primate species. To best reflect this distance, we used the 11,096 one-to-one ortholog alignments to construct a single codon-based reference tree for use in all ML analyses. To ensure we studied only the strongest signatures of positive selection, we analyzed four combinations of evolutionary model parameters and required genes to test significant across all of them ($P < 0.05$, after Benjamini-Hochberg correction for testing 11,096 alignments). Finally, we obtained the set of positively selected codons (and their corresponding amino acid residues) using stringent Bayesian calculations ($P_{posterior} > 0.99$; Methods). These steps resulted in 416 apparent Positive
Selected Genes (aPSG) inferred to contain at least one apparent Positively Selected Residue (aPSR; 1405 in total, Figure S2, Table S3).

Evolutionary models estimate substitution rates in primate protein-coding genes
Our comprehensive phylogenetic tree, calculated for the ML analyses and based on all protein-coding, one-to-one orthologous genes of the nine species studied, has identical topology to the well-supported primate phylogeny (Figure 2, S1)[343]. The five hominoids cluster together, with chimpanzee closest to human followed by gorilla, orangutan and gibbon. Within the Old World Monkeys, macaque and baboon are closest to each other with vervet more distant. Marmoset, the sole high-quality whole-genome sequenced New World monkey, is substantially different from the other species, with a distance to human of 0.15 substitutions per codon compared to ~0.085 between the Old World Monkeys and human. The overall \(d_n/d_s\) rate across our one-to-one protein-coding orthologs is 0.21 (\(d_n = 0.0477, d_s = 0.2235\); Table 1), consistent with average strong purifying

![Codon-based phylogenetic tree of nine simian primates.](image)

**Figure 2. Codon-based phylogenetic tree of nine simian primates.** Branch lengths (nucleotide substitutions per codon) were estimated using the codeml M0 (F61) evolutionary model on the concatenated, masked alignment of 11,096 protein-coding, one-to-one orthologous genes of the nine primates studied. See Figure S1 for species image credits.

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species investigated</td>
<td>9</td>
<td>human, chimpanzee, gorilla, orangutan, gibbon, macaque, baboon, vervet, marmoset</td>
</tr>
<tr>
<td>One-to-one ortholog clusters analyzed for positive selection</td>
<td>11,096</td>
<td>50% of human protein-coding genes; 6.6 million codons human; 57.4 million codons total</td>
</tr>
<tr>
<td>Overall (d_n/d_s) rate (across 11,096 clusters)</td>
<td>0.21</td>
<td>(d_n = 0.0477, d_s = 0.2235)</td>
</tr>
<tr>
<td>Positively selected genes (PSG) (^a)</td>
<td>331</td>
<td>3% of genes tested</td>
</tr>
<tr>
<td>Positively selected residues (PSR)</td>
<td>934</td>
<td>0.5% of PSG codons</td>
</tr>
<tr>
<td>PSG (d_n/d_s) rate (^b)</td>
<td>6.85 (median)</td>
<td>2.10 (minimum)</td>
</tr>
</tbody>
</table>

\(^a\) Genes for which across all four evolutionary models tested: (i) the selection model gives a significantly better fit than the neutral model and (ii) that contain at least one significant PSR (Methods).

\(^b\) Calculated from the per-PSG averages of the \(d_n/d_s\) rates estimated in the four evolutionary models tested (Methods).
selection. Compared to other phylogenetic trees, the estimated substitution rates per nucleotide follow the expected pattern, depending on the fraction of noncoding sequences used for tree reconstruction: the nucleotide-converted branch lengths of our coding sequence tree are a factor 0.87 shorter (median of all branches) than those of a phylogeny based on genomic regions of 54 primate genes (consisting half of noncoding segments) (343), and a factor 0.46 shorter than a tree based on whole-genome alignment (27) (Figure S1).

Rigorous quality control reveals artefacts in 20% of apparent PSG

To assess the reliability of our pipeline and estimate the impact of common errors in large-scale sequence analysis, we performed systematic inspection of all 1405 aPSR and 416 aPSG and implemented filters for automatic detection of false predictions of positive selection (referred to as artefacts; Methods). Although the large majority of predictions are reliable, 85 aPSG (20%) contain artefacts.

The artefacts we encountered fall into five classes (Figure 3A, Table S3). (I) Orthology inference or gene clustering (occurring in 8 of 85 problem cases, 9%). One gene cluster consists of the TRIM60 sequences of seven primates plus the TRIM75 sequences of baboon and gibbon. TRIM60 and TRIM75 are close paralogs (46% identical amino acids, Figure 3B) encoded within a 43kb region on chromosome 4 in human. TRIM75P is an annotated pseudogene in human and TRIM60P is a predicted pseudogene in gibbon, despite both appearing like fully functional genes: they lack frame-shift mutations or premature stop codons and remain highly similar to the corresponding non-pseudogenic copy in the other species (97-98% identical amino acids). Clustering of the outparalogs TRIM60 and TRIM75 led to artificially high substitution rates and an abundance of apparent PSR across the full alignment (Figure S3). (II) Alternative transcript definitions (52 cases, 61%). In most cases this leads to alignment of divergent exons (Figure S3). The CALU gene cluster consists of the coding sequences from two alternative isoforms that include either one of two genomic neighboring and homologous exons (Figure 3C). Given their strong sequence conservation (64% identical amino acids), these mutually exclusive exons likely originated from a tandem exon duplication event (347). For six primates the CALU gene model predicted a single transcript isoform that contains the first of the tandem exons, which made these sequences inconsistent with the alternative transcript selected for the other primates. (III) Unreliable N-/C-termini (32 cases, 38%). When they are sufficiently similar, alternative non-homologous translation starts/stops or alternative exon boundaries may still be aligned. Such cases appear as divergent, high $d_\text{N}/d_\text{S}$ codons (Figure S3). (IV) Alignment ambiguities (4 cases, 5%). These are spurious cases where short sequence regions are hard to align, usually surrounding residues masked for low alignment quality (Table S3). Without independent data (e.g. more sequences, aligned structures) it remains challenging to determine the correct alignment. (V) Other issues (3 cases, 4%). These cases involve apparent PSR with high posterior probabilities in all ML model parameter combinations, but upon closer inspection are untrustworthy (Table S3). One site consisted of distinct codons distributed across the tree (6x AGC, 3x TCA), which were inferred to evolve under high $d_\text{N}/d_\text{S}$ despite all encoding serine residues.
GC-biased gene conversion does not systematically affect the PSG

In addition to positive selection, GC-biased gene conversion (gBGC) may be an alternative explanation for the accelerated evolution of some PSG [348]. gBGC leads to increased GC content in meiotic recombination hotspots, which may inflate $d_\text{w}/d_\text{s}$ estimates [349]. Genes affected by gBGC are expected to have stronger GC substitution patterns than genes evolving under positive selection, particularly in the selectively neutral fourfold degenerate (FFD) sites. Our PSG have significantly lower rather than higher GC contents compared to non-PSG in all studied primates, both across the full coding sequence as well as in FFD sites (human: median GC$_\text{FFD}$ = 0.518 (PSG), 0.578 (non-PSG), $P = 5.3 \times 10^{-6}$, Mann-Whitney $U$ test; Figure S4). Amino acids enriched for PSR positions also correlate only marginally with GC content (Pearson’s $R = 0.14$, $P = 0.56$; Figure S5). PSG and non-PSG overlap to similar extent with female and male human recombination hotspots (7-8%), as well as with genomic regions predicted to be affected by gBGC [350](9% of PSG vs. 11% of...
non-PSG). Furthermore, our site-specific PSG inferences are based on substitution patterns across a primate phylogeny covering at least 36 million years of evolution, which far outdates the rapid turnover rate of recombination hotspots (e.g. even human and chimp hotspot rarely overlap [351]), strongly reducing the influence of gBGC [349]. Together these data suggest that although gBGC may affect individual PSG, the adaptive signatures in the large majority of PSG are not caused by gBGC but are likely the result of positive selection.

Strong statistical evidence for positive selection in 3% of protein coding genes
Removal of the 85 classified artefacts resulted in a final curated dataset of 331 human genes with extensive statistical evidence for positive selection across nine primates (331 PSG, 3% of 11,096 ortholog clusters analyzed; Table 1, S4). The PSG are distributed evenly across the human genome (Figure S6). They have a median \( d_N/d_S \) rate of 6.85 (minimum 2.10), consistent with the strong positive selection observed across multiple evolutionary models. The 331 PSG contain 934 positively selected codons (934 PSR, 0.5% of all PSG codons, average of 2.8 PSR per PSG; Table 1, S5). Over half (53%) of PSG have a single PSR and 14 genes have 12 or more PSR with a maximum of 38: from high to low, MUC13, PASD1, NAPSA, PTPRC, APOL6, MS4A12, CD59, SCGB1D2, PIP, CFH, RARRES3, OAS1, TSPAN8, TRIMS. We observed a notable enrichment of basic residues arginine and histidine among the 934 PSR (Figure S7), which may indicate protein interaction functionality [352].

Immune pathways and functions are abundant among PSG
To gain more insight into the processes that evolved under positive selective pressure in primates, we investigated our 331 PSG for various properties, molecular functions and pathways. PSG are strongly enriched for a variety of immune-related pathways and gene ontology categories of both innate and adaptive nature, including functions in: inflammation, complement cascades, hematopoiesis, B- and T-cell immunity and the defense response against bacteria and viruses (Table S6, S7). Overlap analysis confirmed the enrichment of innate immune functions among PSG (49/331 genes [15%], 2.8-fold enriched compared to non-PSG, \( P = 2.8 \times 10^{-10} \), two-tailed Fisher’s exact test) and pattern recognition pathways components (12 genes, Table S8). These include TLR1, TLR8, MAVS, IFI16, CASP10, TRIMS, OAS1, RNASEL, PGLYRP1, NLRP11, CLEC1A and CLEC4A (Table 2). In addition, many PSG encode transmembrane (65 genes [20%], 1.4-fold, \( P = 6.4 \times 10^{-3} \)) or secreted proteins (73 genes [22%], 2.5-fold, \( P = 1.4 \times 10^{-12} \)), as also indicated by the abundance of enriched terms associated with: extracellular and cell surface localization, receptor activity, signal peptide, disulfide bond and glycosylation (Table S6, S7).

Detailed examination revealed a range of noteworthy genes and processes among the PSG, including: cytokines and their receptors (e.g. IL3, IL4R and CASP1, which activates IL-1β and IL-18), various immunological marker molecules (20 ‘cluster of differentiation’ genes, including CD4/5/48 and the sialic acid binding Ig-like lectins SIGLEC2/3/6), an MHC class II subunit (HLA-DPA1), genes with antimicrobial activity (defensins, granzymes), olfactory and taste receptors (IR10Q1, TAS2R20), ion channels (solute carrier, chloride, sodium channel families), nuclear genes encoding mitochondrial proteins (19 genes, e.g. OXPHOS complex I subunit NDUF10 and assembly factor TMEM126B), a keratin associated protein (KRTAP24-1), poly (ADP-ribose) polymerases (PARP9/13/14), apolipoproteins (APOD/L6/B), and various genes of unknown function (~20-25% of 331 PSG; Table 2, S8). Interestingly, we also find considerable evidence for positive selection associated with
reproduction: genes involved in spermatogenesis and expressed in the testis (e.g. TEX11, CATSPER1, SPATA32), and genes involved in the chromatin structure of the centromere, the kinetochore, chromosome segregation and meiosis (e.g. CENPO/T, REC8, PCNT; see Discussion).

**Positive selection identifies known and novel virus-host genetic conflicts**

The strong signal for immunity among the positive selection data suggests that at least some rapidly evolving sites are in a genetic conflict with one or more pathogens. To further assess the ability of the PSG data to detect such conflicts, we investigated our dataset for known virus-human evolutionary interactions. For that, we evaluated five cases with experimental evidence supporting the importance of positively selected codons (identified through evolutionary analysis) for the ability to restrict infection. Our large-scale approach correctly identified four out of five cases (TRIM5, MAVS, BST2 [tetherin], SAMHD1 [71,211,353,354]; Table S9). Moreover, despite using sequences from substantially less species (those case studies sequenced one gene in ~20-30 primates), we retrieved many of the reported codons (7/12 in TRIM5, 1/1 in MAVS, 0/1 in BST2, 1/3 in SAMHD1). In addition to case studies with both statistical and experimental support, our large-scale automated approach also recovered antiviral PSG for which just statistical evidence for positive selection was reported in small-scale studies (IFI16, OAS1, TLR8, PARP9/13/14, APOL6 [355-359]; Table S9).

Next, we probed our evolutionary data for novel cases of virus-human genetic conflicts. To prioritize the PSR and PSG, we integrated them with a multitude of orthogonal datasets describing various aspects of antiviral immunity and the virus-host interaction (Table S8), including virus-human protein interactions [33], functional screens of virus infection [34], gene expression in immune cell subtypes [84] and virus-infected cells [326], and maps of recent human adaptation [360]. PSG are

<table>
<thead>
<tr>
<th>Biological process or gene family</th>
<th>Selected genes 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern recognition &amp; defense pathways</td>
<td>TLR1/8, MAVS, IFI16, CASP10, TRIM5, OAS1, RNASEL, PGLYRP1, NLRP11, CLEC1A/4A</td>
</tr>
<tr>
<td>Cytokines &amp; receptors</td>
<td>IL3/13/25, IL4R, CXCR/R2/L16, CCRL2, CASP1</td>
</tr>
<tr>
<td>Immunological (marker) molecules</td>
<td>20 CD genes; CD4/5/1/1C/4B/58/…; CD22 (SIGLEC2), CD33 (SIGLEC3), CD33L (SIGLEC6), HLA-DPA1</td>
</tr>
<tr>
<td>Complement system</td>
<td>CS/BB/9, CR2, CFH, C4D/65/59</td>
</tr>
<tr>
<td>Antimicrobial peptides &amp; cytotoxic activities</td>
<td>DEF61/118/132/136, GZMA/B/K, PRG2/3</td>
</tr>
<tr>
<td>Reproduction &amp; testis-expressed genes</td>
<td>TEX11/26/29, CATSPER1/3/G, SHBG, SPACA7, CYLC1/2, CT55, PATE1, SPATA18/32</td>
</tr>
<tr>
<td>Chromosome segregation &amp; centromeres</td>
<td>CENPO/T, CEP250, PCNT, REC8, SGOL2, MISP, CDKSRAP2</td>
</tr>
<tr>
<td>Smell &amp; taste genes</td>
<td>OR10Q1/G/D3, BPIFB3, TAS2R10/20/41</td>
</tr>
<tr>
<td>Ion channels</td>
<td>SLC6A16/9C1/26A3/44A2, CLCA1/4, SCNN1D/7A/11A</td>
</tr>
<tr>
<td>Mitochondrial proteins</td>
<td>19 genes; NDUFA10, COA1, MRPS30, MTIF3, TMEM126A/B, LRPPRC, TFB2M, …</td>
</tr>
<tr>
<td>Other</td>
<td>KRTAP24-1, HDAC1, APOD/L6/BR, PARP9/13/14</td>
</tr>
<tr>
<td>Uncharacterized genes</td>
<td>CCDC27/57, KIAA1328/0226-like/1407, PRR14/30, SSMEM1, SUN5, TMEM186/215/225, MFSDF7, FAM162A/220A, RNF213, TSPAN8, LLZP4, MS4A12, CMTM6, C5orf45, C3orf17, C1orf76, C1orf168, C20orf96, CXorf66, …</td>
</tr>
</tbody>
</table>

1 See Table S4 for the complete list of PSG.
enriched for genes showing differential expression upon infection with respiratory viruses (12 genes [4%], 3.3-fold, \( P = 5.4 \times 10^{-4} \); Table S8). Many PSG are also involved in virus-human PPIs (70 genes [21%]; though not more often than non-PSG). Among the 70 PSG whose protein products interact with viruses, 49 are expressed in several or all 14 profiled immune cell subtypes (including B-, T-, NK-cells, DCs and monocytes), 11 of which also upon their knockdown alter the course of cellular infection (e.g. with Hepatitis C virus [HCV], Sendai virus [SeV] or human papilloma virus 16 [HPV16]; Table S8). Besides well-described viral interactors such as MAVS, SAMHD1 and CD55, those 11 PSG include poorly characterized genes that may represent novel candidate virus-host interactors. For example, FBXO22 encodes an F-box family protein, which may be involved in ubiquitin-mediated protein degradation. FBXO22 has not been linked to (viral) infections, other than a role in macrophage NFκB activation during Salmonella infection [361]. Our evolutionary data together with the virus-host interaction data suggest a role for FBXO22 in primate HPV infections, possibly mediated by the positively selected position 3 (Pro in human, chimpanzee, baboon, marmoset – Thr, Ser, Ala or Leu in gorilla, orangutan, gibbon, macaque, vervet).

20 other poorly annotated PSG are candidates for virus-host genetic conflicts (Table 2, S8). These include KIAA0586, which shows immune cell expression, has been found to interact with Influenza A virus, and is involved in cilia deficiencies [362] FAM162A interacts with Measles virus and is a likely component of the mitochondrion [189]. Ribosomal protein S29 (RPS29) is not only a strong target of positive selection in our study but it also localizes to a genomic locus that underwent recent human adaptation [360]. RPS29 is expressed in 14 immune cell subsets, interacts with six different Influenza proteins, and affects HCV and SeV replication. Thus, data-driven prioritization of genes under positive selection may reveal novel cases of virus-host interactions shaped by evolutionary conflicts.

**PSR are at the structural interface between viruses and their cellular receptors**

To gain further insights into the role of positive selection in viral infection, we analyzed the involvement of PSR in structurally characterized virus-host interactions. Among the PSG with strong adaptive signatures, we identified three genes that function as virus receptors and for which human-virus protein complexes have been solved: CD4, CD46 and CD55. CD4 interacts with MHC class II molecules and is the classical surface marker of T helper cells (CD4+ T cells). HIV exploits CD4 as a receptor for entering host T-cells [363]. The two residues identified as positively selected by our approach, Asn77 and Ala80, are part of the CD4 V-set Ig-like domain. Both lie close to the interaction interface between CD4 and HIV envelop protein gp120, with CD4 Asn77 making extensive contact and hydrogen bonding to gp120 Ser365 (Figure 4A).

CD55 [DAF] and CD46 [MCP] are members of the regulator of complement activation (RCA) gene family that control activation of the complement cascade (21% identical amino acids; both composed of four SCR domains). CD55 acts as an entry receptor for some picornaviruses, including several types of coxsackie-, entero- and echovirus [364,365]. CD55 SNPs are associated with geographical pathogen richness and CD55 in human likely evolved under balancing selection, which maintains allelic diversity in a population [366]. We identified nine primate positively selected positions in CD55 (two in SCR domain 1, two in SCR2, five in SCR3). Arg170 and Gly178 (part of SCR3) are involved in the interaction with echoviruses (EV) 7 and 12, with Arg170 buried deep
into the EV7 viral capsid (Figure 4B). A previous study reported additional interactions between picornaviruses and our CD55 PSR, including Val155 (EV7), Val124 and Ile206 (coxsackievirus B3) [365].

CD46 is a receptor for measles virus (MV), human herpesvirus 6 and several adenovirus subspecies (Av) [367,368]. Protein structure analysis revealed that the single PSR in CD46, Arg103 (SCR domain 2), makes extensive contacts with both MV hemagglutinin (His448, Asn449, Glu471; Figure 4C) and the Av type 21 fiber knob (Asn304, Ile305). Together, these results suggest that positions evolving under strong positive selection may be central to the interactions between virus particles and their human cellular receptors.

**Discussion**

In this study we presented an unbiased screen to identify novel virus-host genetic conflicts on the basis of statistical evolutionary evidence for molecular adaptation between species. While previous investigations of virus-host genetic conflicts have been limited to single genes [71-73,211] or subsets...
of genes known to be important for infection [369,370], we took a generic approach starting from entire genomes. This was enabled by the recent completion of several primate genome sequences bridging key evolutionary timescales between previously available species.

Earlier estimates of primate genes undergoing positive selection range from ~1% to 4-6% to ~10% depending on the species studied, the detection method and the number of genes tested [62-64,68,333,334,336,371]. Lower estimates were likely underpowered and limited by genome availability (e.g., comparing human-chimp), while higher estimates tend to be affected by the high rate of false positive predictions discussed before [46,336-341]. Our observed $d_n/d_s$ rate across one-to-one protein-coding orthologs (0.21) is similar though slightly smaller than previous estimates based on human, chimp, and macaque (0.23-0.26 [64,333,371]). Given our conservative approach (minimizing false positives while perhaps missing potentially interesting sites) applied to half of all human protein-coding genes, our estimate that 3% of genes are under positive selection should represent a reliable lower limit detectable using the current whole-genome sequenced primates.

The strict nature of our pipeline suggests there still exists a large discrepancy between data availability and the reliability of large-scale comparative sequence analysis. Despite their limited evolutionary distance, even the study of primate genomes encounters many challenges. The large majority of positive selection artefacts remaining after stringent automated filtering arise from alignment of coding sequences that are not strictly from the homologous genomic region in different species, i.e., they arise from alignment of non-orthologous codons. The predominant underlying source of these issues are inconsistencies in gene models and genome annotation, which causes differences in coding sequence start/stop locations, exon boundaries, pseudogene predictions and alternative transcripts. Thus, rigorous manual inspection and curation at all stages of automated pipelines remain critical for reliable results and provide insights into the current challenges in comparative genomics studies.

Genomes evolve under a huge variety of selective pressures, only some of which are caused by pathogens. On top of this, each complete species is represented by a single consensus sequence. Our investigation of primate genomes detected traces of genetic conflicts driven by exposure to historical viruses, some of which may also influence susceptibility to present-day viruses. Complementary studies of genetic variation within current human populations and archaic hominins (e.g., Neanderthals, Denisovans) may identify targets of positive selection caused by more recent selective pressures [59,360]. For example, human SNPs that correlate with current geographic virus diversity were shown to be enriched for immune genes [372]. Stronger deduction of the selective forces, such as pathogens, driving genetic variation would require investigations of a continuum of individuals, populations and species.

Selection does not usually operate on the level of single genes. It operates within the context of an entire genome, which consists of two alleles for most genes and is the result of a complex evolutionary history involving many (partial) gene duplications and constant genetic recombination. Such factors enable genetic diversification and may account for the ability of animal genomes to sustain genetic conflict with the incredible diversity of rapidly mutating, highly abundant viruses [69]. Thus, aside from variation on the codon level in one-to-one orthologs, duplications of genes
and exons (some of which we report in Table S3) may represent other evolutionary mechanisms facilitating (antiviral) adaptation.

Our positively selection approach successfully identifies genes with a role in immunity and offer insights into other functions and cellular systems that are evolving adaptively. A striking number of positively selected genes (PSG) act as cell surface receptors in adaptive immunity. Others are involved in cytokine signaling and innate immune functions such as the complement system, intracellular pathogen recognition (both receptors and pathway members) or intrinsic antiviral activity. We also found adaptive signatures potentially related to other phenotypes that may be of interest, including morphology (hair protein KRTAP24-1), dietary diversity (smell and taste receptors), cholesterol and lipid transport (apolipoproteins), and energy metabolism (mitochondrial proteins).

Another exciting group of PSG are those involved in centromere structure, chromosome segregation and meiosis, which may be related to the phenomenon of centromere meiotic drive [373]. Meiosis in female animals is asymmetric in that only one haploid gamete is retained. Retention of individual chromosomes depends on a preferred binding orientation of the centromeres to the microtubular spindle via the kinetochore. Centromere evolution has been theorized to be under strong Darwinian selection because ‘selfish’ centromeres with a favored retention in meiosis are directly more likely to be transmitted [373]. This process may have driven the evolution of the highly variable centromere DNA sequence, marked by a specific chromatin structure containing the CENPA histone H3 variant [373,374]. We found positive selection signatures in the centromere-associated proteins CENPT and CENPO, both involved in mediating the interaction between CENPA nucleosomes and the kinetochore. In addition, our PSG contain two genes involved in the meiotic cohesion complex (REC8, SGOL2) and various genes involved in the centrosome and spindle machinery (CEP250, PCNT, CDK5RAP2, MISP). The rapid evolution of the centromere may be a driver contributing to the observed adaptations in these genes, even though a direct interaction with the centromere has not been established in all cases.

Sites under positive selection are probably important for the biological processes that led to their selection, for instance through physical interactions (e.g. between viral and human proteins). However, deconvoluting the contributions of specific selective pressures to the complex landscape of genome variation requires additional information. Building on this concept, we integrated the genomic signals of positive selection with orthogonal data describing the virus-host interaction to not only predict novel virus-host genetic conflicts, but also to identify positions likely important in these interactions (Table S8). Protein structure analyses revealed close contacts between virus particles and positively selected residues in cellular entry receptors for HIV, measles, adenoviruses and picornaviruses (CD4, CD46, CD55). These observations, guided by systematic analysis of virus-host evolutionary genetics, suggest positions in both viral proteins and their human receptors that are important for infection and may represent novel targets for vaccine or antibody development.
### Methods

#### Sequences
We obtained protein-coding DNA sequences of all nine simian primates for which high-coverage whole-genome sequences are currently available from Ensembl release 78, December 2014 [27] (Table S1, Figure S1). We processed orthology definitions from the Ensembl Compara pipeline [43] to obtain 11,170 one-to-one ortholog clusters containing for all nine primates a single coding sequence corresponding to the canonical transcript, which usually encodes the longest translated protein (Table S2). Clusters consist of genes for which only one copy is found in each species, and these genes are one-to-one orthologs to the human gene. Sequences are of high quality as indicated by the general lack of undetermined nucleotides ('N'): 98,420 of 100,530 sequences (98%) and 9,312 of 11,170 (83%) clusters contain no Ns. Genomes with most Ns are gibbon (in 680 sequences), gorilla (313) and marmoset (291).

#### Initial alignments
We first obtained multiple alignments of the clusters of orthologous primate protein sequences using MUSCLE [375] and Mafft [376], and from the Compara pipeline (i.e. filtering the larger vertebrate alignment for primate sequences) [43]. Inspections revealed that misalignment of nonhomologous codons affects virtually all alignments, as was observed in previous studies [336-340]. This is probably the result of the tendency of alignment algorithms to overalign, i.e. produce alignments that are shorter than the true solution due to collapsed insertions in an attempt to avoid gap penalties [337,344]. The PRANK algorithm to an extent prevents alignment of nonhomologous regions by flagging gaps made during different stages of progressive alignment and permitting their reuse without further penalties [344]. As PRANK has been shown to provide better initial alignments for large-scale positive selection detection [336-341], we obtained multiple alignments of the primate ortholog clusters using the PRANK codon mode (prank +F –codon; v.140603). We used the default settings of (i) obtaining a guide tree from MAFFT for the progressive alignment procedure and (ii) selecting the best alignment from five iterations. These settings likely result in the best alignment for a given cluster of sequences (including those showing a gene tree topology that differs from the species tree topology). The PRANK approach markedly improved the initial alignments.

#### Alignment filtering and masking
Even with improved initial alignments, positive selection studies remain affected by a high rate of false positive predictions. Part of those may be alleviated by additional automated masking of unreliable alignment regions. GUIDANCE assesses the sensitivity of the alignment to perturbations of the guide tree [345] and has been recommended for positive selection studies [338,339,341]. We applied GUIDANCE with the default 100 bootstrap tree iterations (guidance.pl --program GUIDANCE --seqType nuc --msaProgram PRANK --MSA_Param “\+F \-codon”; v1.5). TCS assesses alignment stability by independently re-aligning all possible pairs of sequences and scoring positions through comparison with the multiple alignment [346]. We ran TCS on translated PRANK codon alignments (t_coffee -other_pg seq_reformat -action
Low confidence scores of either method led us to remove entire alignments from our analysis and mask individual columns and codons. Alignments were removed in the case of a low score (default cutoffs of <60% for GUIDANCE, <50% for TCS) for (i) the overall alignment or (ii) one or more sequences (i.e. we only retained alignments with sequences for all nine species). Entire columns were masked if GUIDANCE <93% or TCS <4; individual codons were masked if <90% or <4. Masked nucleotides were converted to ‘n’ characters to distinguish them from undetermined nucleotides in the genome assemblies (‘N’). For visualization and quality inspection purposes we translated the masked codon alignments to the corresponding protein alignment. Nucleotides ‘n’ and ‘N’ were converted to ‘o’ and ‘X’ upon translation, respectively. Detailed visual inspection revealed the value of our masking approach: masked codons tend to comprise unreliable alignment regions, primarily consisting of large inserts, insertion-deletion boundaries (i.e. regions bordering well-aligned blocks), and aligned but nonhomologous codons.

Evolutionary analyses: reference phylogenetic tree
Maximum likelihood (ML) $d_N/d_S$ analysis to infer positive selection of genes and codons was performed with codeml of the PAML software package v4.8a [61] (Text S1). We used a single phylogenetic tree with branch lengths for the ML analysis of all alignments to limit the influence of gene-specific phylogenetic variability. To obtain this reference tree, which could not be obtained from published work, we concatenated all 11,096 masked alignments into one large alignment and ran the codeml M0 model (i.e. fitting a single $d_N/d_S$ for all sites; NSsites = 0, model = 0, method = 1, fix_blength = 0), provided with the well-supported topology of the primate phylogeny [27,343]. We took this approach for two main reasons: (i) to best reflect the overall evolutionary distance between the primate species (which influences codon transition probabilities in the ML calculations, Text S1), and (ii) to estimate branch lengths in units compatible with codon-based evolutionary analyses, i.e. the number of nucleotide substitutions per codon. For comparisons with other primate phylogenetic trees, the branch lengths of our codon-based tree were converted to nucleotide substitutions per site (i.e. nucleotide substitutions per codon divided by three). The codeml M0 model under the F61 or F3X4 codon frequency parameters resulted in virtually identical phylogenetic trees (median branch length difference of a factor 0.99) and $d_N/d_S$ estimates (0.213 vs. 0.217; Figure S1). The M0 tree is also highly similar to a ML phylogenetic tree inferred from the same concatenated alignment using nucleotide rather than codon substitution evolutionary models (median branch length difference of a factor 0.98; Figure S1; RAxML v7.2.8a [377]; -f a -m GTRCAT -N 1 0 0).

Evolutionary analyses: inference of positive selection
In the first of two steps for inferring positive selection using codeml, the 11,096 filtered and masked alignments were subjected to ML analysis under evolutionary models that limit $d_N/d_S$ to range from 0 to 1 (neutral model) and under models that allow $d_N/d_S > 1$ (selection model; Text S1) [60]. Genes were inferred to have evolved under positive selection if the likelihood ratio test (LRT) indicates that the selection model provides a significantly better fit to the data than does the neutral model ($P_{LRT} < 0.05$, after Benjamini Hochberg correction for testing 11,096 genes). We included apparent
Positively Selected Genes (aPSG) if they met the LRT significance criteria under all four tested ML parameter combinations. These combinations consist of two sets of evolutionary models: M1a (neutral) vs. M2a (selection); M7 (beta) vs. M8 (beta&ω). And two codon frequency models: F61 (empirical estimates for the frequency of each codon); F3X4 (calculated from the average nucleotide frequencies at the three codon positions). I.e. we used combinations of the following codeml parameters: NSsites = 1 2 or NSsites = 7 8; CodonFreq = 2 or CodonFreq = 3; cleandata = 0, method = 0, fix_blength = 2. 2,992 (27%) genes showed significant evidence of apparent positive selection at the level of the whole alignment (Figure S2A).

Second, for the significant aPSG we retrieved from the site-specific codeml ML analyses (step one, above) the Bayesian posterior probabilities indicating which individual codons may have evolved under positive selection (Text S1). We included apparent Positively Selected Residues (aPSR) if their codons were assigned high posteriors under all four ML parameter combinations (P_posterior (ω > 1) > 0.99). 416 aPSG contain at least one significant aPSR (1405 in total; Figure S2B).

Quality control
We subjected each inferred aPSR and aPSG to visual inspection (Table S3). In this way we identified several indicators for positive selection artefacts that we then used for their automated detection. First, we obtained the gene trees for our individual masked alignments using RAxML [-f a -m GTRGAMMAI -N 100]. Type-I [orthology] and -II [transcript definitions] artefacts tend to lead to gene trees with (i) a long-branched clade consisting of the set of sequences that are distinct from the others (e.g. paralogs, alternative exons), and (ii) a topology that is not congruent with the well-supported species phylogeny (Figure S3). We filtered out likely false positives by looking for gene trees with an extreme longest/average branch length ratio. Second, to assess the distribution of PSR across exons, we mapped Ensembl exon coordinates for human transcripts to the human protein sequences. Type-II [transcript definitions] and -III [termini] artefacts could often be filtered out by a high concentration of aPSR located to a single exon.

GC-biased gene conversion (gBGC)
The effects of gBGC seem specifically correlated to regions of high meiotic recombination in males rather than females. We calculated genomic overlaps of PSG and non-PSG with male (8.2% of PSG, 7.7% of non-PSG) and female (6.7% of PSG, 8.1% of non-PSG) recombination hotspots in human, which we obtained from the family-based deCODE maps via the UCSC genome browser. Sex-averaged recombination hotspots estimated from linkage disequilibrium patterns were obtained from HapMap Release 22 (43% of PSG, 39% of non-PSG). Human genomic regions under the influence of gBGC were predicted by phastBias (9.1% of PSG, 11.4% of non-PSG).

PSG function analyses
Our final curated set of 331 PSG (Table S4, S8) was analyzed for gene ontology terms, pathways databases and other functions (Table S6, S7), compared to a background of 11,011 genes (the 11,096 genes tested for positive selection excluding the 85 artefacts). Enrichment statistics were calculated for overlaps between the PSG and various gene lists (Table S8), assessing whether a gene list of interest was significantly over-represented among the 331 PSG compared to the 10,680 non-PSG (two-tailed Fisher’s exact test on a 2x2 contingency table of the overlap). Secreted
(keyword KW-0964) and membrane (KW-0472) proteins were obtained from UniProt [244]. Innate immunity, PRR and virus-human PPI gene lists are described in [326]. We mined the GenomeRNAi database [34] for genes whose perturbation significantly affects viral infection or replication. Data on gene expression in mouse immune cell subsets were obtained from the Immunological Genome Project (using the recommended expression thresholds – usually 120) [84] and mapped to human orthologs. Optimized atomic-resolution protein structures of virus-host interactions were obtained from PDB_REDO [380], visualized with YASARA (www.yasara.org), and analyzed using WHAT IF [381].

Scripts and tools
Our pipeline consists of custom Perl and R scripts, available upon request. In addition to the methods cited above, we made extensive use of the Ensembl API version 78 [27], GNU Parallel [382] and Jalview [383].

Acknowledgments
We acknowledge Manja Leemans for exploratory work on the applied methods. We also thank Dei M. Elurbe, Colin Logie, Eelco Tromer, Geert Kops and Berend Snel for stimulating discussions. This work was supported by the Virgo consortium, funded by the Dutch government (FES0908), and by the Netherlands Genomics Initiative (050-060-452).
Supplementary Text

Text S1. Introduction to molecular adaptation and inference of positive selection from comparative evolutionary analyses.

**Molecular adaptation caused by positive selection**

Acquired phenotypic traits can have three broad types of effects: (i) the trait has no effect on the ability to survive and reproduce (fitness) of the individual carrying it (i.e. the trait will be selectively neutral), (ii) the trait has a negative effect on fitness and its prevalence in the population should decrease (a process referred to as negative or purifying selection), or (iii) the trait has a positive effect on fitness and its prevalence in the population should increase (positive Darwinian selection). The actual ability of natural selection [56] to change the prevalence of a phenotypic trait in a population depends on a variety of factors, including the magnitude of the fitness advantage or disadvantage caused by the trait and the effective size of the population.

The neutral theory of molecular evolution [384] proposes that the majority of evolutionary changes at the molecular level are deleterious and therefore purged by negative selection, while most of the remaining mutations are neutral rather than beneficial. According to this view, beneficial mutations leading to positive selection are rare. Therefore, the majority of observed molecular variation in populations has been proposed to be the result of random sampling of selectively neutral or nearly neutral genetic variants (genetic drift). However, despite their proposed rarity, mutations creating advantageous phenotypes can be a great source of insights into the adaptive evolutionary processes governing the development of species and populations.

Changes in the DNA of a species (or other heritable features, e.g. epigenetic changes) can result in variety of molecular traits that might influence cellular and organismal phenotypes, and hence fitness (though the contribution of individual variations to fitness tends to be small [385]). For instance, mutations in the regulatory elements of the genome influence transcription and gene expression, while mutations to the coding sequence of genes (cDNA) can lead to changes in the protein structure. Since a large part of cellular functionality is achieved through proteins, such protein-changing mutations can have direct consequences for cellular phenotype. Therefore, many studies on molecular adaptation focus on the protein-coding regions of the genome. Statistical methods have been developed (i) for detecting which genes are likely to have evolved under positive selection, (ii) for inferring which codons within those genes are responsible for their positive selection signatures, and (iii) for finding out whether specific lineages or sequences underwent episodes of adaptive change [386].

**Inferring genes and codons under positive selection by \( d_{\text{NS}}/d_{\text{S}} \)-based maximum likelihood methods**

Because of the degeneracy in the genetic code (i.e. all amino acids except methionine and tryptophan can be encoded by multiple codons), some mutations in the cDNA cause changes to the protein sequence (nonsynonymous substitutions), while other mutations do not change the protein sequence (silent, synonymous substitutions). An excess of nonsynonymous substitutions (normalized by the total possible nonsynonymous sites, \( d_{\text{NS}} \)) over synonymous substitutions
Positive selection in the human-virus interaction

(normalized by the total possible synonymous sites, \(d_S\)) has proved to be a good indicator of positive selection acting on a gene \([60]\). Interpretation of this \(\omega = d_N/d_S\) ratio (also known as \(K_a/K_s\)) is based on the assumption that silent substitutions are neutral and fixed by random genetic drift. Therefore, \(d_S\) can serve as a baseline for the expected rate of (neutral) nonsynonymous substitution. Thus, the mode of selection acting on a gene or a specific codon can be inferred from the value for \(\omega\) as follows:

- \(\omega = 1\); equal rates of \(d_N\) and \(d_S\); nonsynonymous substitutions are neutral on average
- \(\omega < 1\); lower rates of \(d_N\) than \(d_S\); part of the nonsynonymous substitutions are deleterious and purged by negative/purifying selection
- \(\omega > 1\); higher rates of \(d_N\) than \(d_S\); part of the nonsynonymous substitutions are beneficial and retained by positive selection

It should be noted that estimates of \(\omega\) for an entire gene represents an average of the \(d_N/d_S\) ratios of all codons, which are likely to be heterogeneous and many of which are probably strongly conserved. Thus, \(d_N/d_S\)-based methods tend to require a strong signal across multiple lineages in order to identify positive selection.

Several evolutionary phenomena and assumptions confound \(d_N/d_S\) estimates. First, synonymous substitutions are not always neutral as they can influence for example mRNA stability and splicing, and can even influence the final protein, perhaps by influencing co-translational folding \([387]\). Furthermore, (i) nucleotide substitutions show a higher rate of transitions than transversions, (ii) different codons encoding the same amino acid are not used in equal amounts (biased codon usage), (iii) multiple substitutions may have occurred over time at the same position, other than the differences present between the final observed sequences, and (iv) different sequences (species) have undergone different divergence times between them, which influences expected amounts of variation between sequences \([60,388]\). Analyses of \(d_N/d_S\) work best for sequences that are sampled across an appropriate, ‘medium’ evolutionary distance: they should be distant enough to show some variation, but not too divergent to limit the effects of multiple substitutions occurring at the same site and prevent the associated loss of information \([389]\).

Maximum likelihood (ML) methods have been shown to best account for these confounding factors and achieve reliable \(d_N/d_S\) estimates based on evolutionary models. They work by simultaneously estimating the combination of parameters (the \(d_N/d_S\) ratio \(\omega\), the transition/transversion rate ratio \(\kappa\), codon frequencies \(\pi\)) that best explains the sequence data (in the form of a multiple sequence alignment), given the divergence time \(t\) between sequences (i.e. expected amount of change, as represented by the branch lengths of a phylogenetic tree \([60]\)). The codeml program in the PAML software package (‘Phylogenetic Analysis by Maximum Likelihood’) is most popular \([61]\) for detecting positive selection in multiple sequence alignments of protein coding sequences.

The method explores many combinations of values for parameters \(\omega\), \(\kappa\), and \(\pi\) and for each combination constructs a matrix of codon transition probabilities over time \(t\), \(P(t)\) \([60]\). The value for time \(t\) between two sequences (divergence) can either also be explored as a free parameter, or can be fixed based on external data (branch lengths in a provided phylogenetic tree). The probability of one codon replacing another represents a Markov process of codon substitution \([390]\), which only depends on the current identity of the codons and is independent of their
history. Each codon transition probability state is then used to calculate the likelihood that the corresponding combination of parameters would have given rise to the data in the sequence alignment (i.e. the codons observed across the sequences). For each codon position (referred to as a ‘site’) in the alignment, the likelihood that a particular transition probability matrix gave rise to the observed codon configuration across the sequences is obtained by multiplying the probabilities of substituting one codon with another, at each part of the tree, summing over all possible ancestral states [391]. Under the assumption that sites evolve independently from each other, the likelihood for the full alignment is obtained by multiplying the likelihoods per site. Maximizing the full-alignment likelihood eventually leads to the combination of parameters, including \( \omega \), \( \kappa \), and \( \pi \), that best fit the data.

Genes can be assessed for different ‘models’ of natural selection by varying the range of values that \( \omega \) is allowed to take in the maximum likelihood calculations. To test for positive selection, one can ask whether a null model (or neutral model; \( \omega \) is only allowed to range from 0 to 1) better fits the data than an alternative model (or selection model; \( \omega \) is allowed to take values larger than 1) [60,335]. The likelihood ratio test assesses whether the alternative model better explains the data than does the null model, by comparing twice the log-likelihood difference between the two models to a chi-square test distribution (LRT = 2x(\ln L[selection model] − \ln L[neutral model]). A statistically significant better fit of the alternative model suggests that positive selection may have acted on the gene.

Various neutral and selection site models have been defined [60,335]. They differ in the distributions from which \( \omega \) is sampled during the maximum likelihood parameter estimation. Two model comparisons have the best power and are commonly performed:

(i) \textit{M1a (neutral) vs. M2a (selection)}
Model M1a fits two \( \omega \) site classes in proportions \( p_0 \) and \( p_1 = 1 - p_0 \) with \( 0 < \omega_0 < 1 \) (purifying selection) and \( \omega_1 = 1 \) (neutral). Model M2a adds a proportion \( p_2 = 1 - p_0 - p_1 \) of sites with \( \omega_2 > 1 \) (positive selection).

(ii) \textit{M7 (beta) vs. M8 (beta&\omega)}
Model M7 fits \( \omega \) between 0 and 1 as a beta distribution, the shape of which (e.g. L-, U-, n-, J-shaped) depends on beta distribution parameters \( p \) and \( q \), which are estimated in the ML calculation. In model M8, a proportion \( p_0 \) of sites have \( \omega \) drawn from a beta distribution and a proportion \( p_1 = 1 - p_0 \) of sites have a single \( \omega_1 > 1 \).

Genes that shown evidence of having evolved under positive selection according to maximum likelihood analysis can then be studied to determine which sites (codons, and thus residues/amino acids) are subject to selection. The ML procedure results in an estimated distribution of \( \omega \) values across the gene as well as the estimated proportions with which these \( \omega \) values occur. Inference of which sites evolved under positive selection follows a Bayesian approach (Bayes empirical Bayes), which uses the ML estimates for the values and proportions of \( \omega \) classes as prior probabilities for each site in the alignment [60,378]. The prior probabilities are updated according to the data at a site in the alignment (the observed codons across the sequences), to arrive at site-specific posterior probabilities for each class of \( \omega \) values: 

\[
P_{\text{posterior}}(\omega \mid \text{data}) = P_{\text{prior}}(\omega) \cdot P(\text{data} \mid \omega) / P(\text{data})
\]

The \( \omega \) class that maximizes the posterior probability is the most likely class for the site. A high posterior
probability (typically >0.90 or >0.95) for the \( \omega > 1 \) class suggests that the site in question evolved under positive selection.

**Detecting positive selection from within-species variation**

Besides between-species sequence variation approaches such as \( d_N/d_S \), within-species variation can also be exploited to study molecular adaptation. Many such intraspecies approaches are based on population genetics statistics around the concept of selective sweeps. This model presumes that recent strongly advantageous mutations become rapidly fixed in a population or species [59,392,393]. Due to genetic hitchhiking, fixation of the mutation would cause genetically linked alleles (which usually locate to the region surrounding the mutation) to also increase in frequency or become fixed, while alleles that are not linked to the mutation will decrease in frequency or be lost. Thus, selective sweeps, compared to neutrality, generally result in a decrease in genetic diversity across a larger genomic region surrounding a favorable mutation. The lack of sequence diversity appears for instance as an excess of rare alleles and increased linkage disequilibrium in the region [394].

A final approach for detecting adaptation at the molecular level makes use of both inter- and intra-species variation. The McDonald-Kreitman test compares levels of variation occurring within species (referred to as polymorphism) to levels of variation between species (referred to as divergence)[395]. This allows an estimation of how much of the variation between species is driven to fixation within species. If all substitutions are neutral, both types of variation should be similar, while an excess of nonsynonymous variation between species compared to within species indicates positive selection [395,396].
Figure S1. Phylogenetic trees of the nine simian primates selected for the analyses. Plotted on top of the well-supported primate topology are branch lengths of five different phylogenetic trees. (M0_F61, M0_F3X4) Protein coding-based reference phylogenetic trees used in all ML analyses. These trees were calculated using the codeml M0 evolutionary model under the F61 (M0_F61, same tree as in Figure 2) or F3X4 (M0_F3X4) codon frequency parameters on a concatenated alignment of 11,096 protein-coding, one-to-one orthologous genes of the nine primates studied. Other statistics: [M0_F61] kappa (ts/tv) = 3.91981, dN/dS = 0.21341, dN = 0.0477, dS = 0.2235; [M0_F3X4] kappa (ts/tv) = 4.15152, dN/dS = 0.2168, dN = 0.0484, dS = 0.2311. (RAxML) Maximum likelihood phylogenetic tree of the same concatenated alignment, inferred using nucleotide rather than codon evolutionary models. (Perelman) Nine primates extracted from a 186-primate phylogeny based on genomic regions of 54 primate genes (consisting half of noncoding parts) from Perelman et al. [343]. (Ensembl) Adapted from the full species tree of Ensembl release 78 (December 2014), which is based on the mammals EPO whole-genome multiple alignment pipeline [27]. Branch lengths are in nucleotide substitutions per site, with ‘sites’ being codons in (M0_F61, M0_F3X4) and nucleotides in (RAxML, Perelman, Ensembl). Species pictures were taken from Ensembl and Table S1.

Figure S2. Overlaps between positive selection predictions from four evolutionary model parameters combinations. Apparent Positively Selected Genes (aPSG, A) and Residues (aPSR, B). Only for significant aPSG did we collect aPSR from the site-specific codeml predictions. See Methods. Venn diagrams created using Venny [397].
Figure S3. Examples of positive selection artefacts. (A) A type-I problem (orthology) involving the clustering of outparalogs TRIM60 and TRIM75. The distinct sets of sequences differ across their whole length, leading to artificially high substitution rates across the whole alignment. See also Figure 3B. (B) A type-II problem (transcript definitions) involving mutually exclusive, tandem duplicated exons in the CALU gene. All aPSR locate to a single mapped exon. See also Figure 3C. (C) A type-II problem (transcript definitions) involving three distinct sets of sequences for the USE1 gene across the primate species. These sequence sets originate from different gene models in the different species, some of which include a small exon (top genome browser screenshot), while others have an extended 3' exon boundary (bottom). All aPSR locate to the same small region in the protein. (D) A type-III problem (unreliable C-terminus) in the DNAJB12 gene. In some species the gene model includes a slightly longer last coding exon, while in others it features a small coding region within the region that is part of the UTR in others. All aPSR locate to the small region at the alignment C-terminus. Panels further show alignments that are codon-based, masked and translated, as well as their gene trees. The lower black blocks under the alignments indicate exon coordinates mapped to the protein alignments, while the black bars above the exon blocks indicate the predicted aPSR. Barcodes indicate the distribution of aPSR across the sequences. UCSC Genome Browser screenshots show BLAT alignments of cDNA sequences of the nine primates (black tracks).
Figure S4. Comparison of the GC contents of PSG (N=331) and non-PSG (N=10,839) across the nine primates studied. GC content was calculated as the fraction of nucleotides that are G or C across the full coding sequences (A) or across all fourfold degenerate (FFD) sites (ACN CCN CGN CTN GCN GTN TCN; B). Note that Y-axes have different limits in (A) and (B). Boxplots in both (A) and (B) include all data points, i.e. there are no outliers. FFD sites tend to have slightly higher GC content than the full coding sequences, but PSG have lower GC content than non-PSG. The general lack of housekeeping functions in our PSG may be the cause of the lower GC content of our PSG compared to other genes (i.e. housekeeping-like genes tend to have high GC content).

Figure S5. Correlation between PSR amino acid enrichment scores and codon GC content. Codon GC content was calculated as the average GC content of all codons for a particular amino acid. Plot shows the linear regression line with 95% CI. See Figure S7 for the PSR enrichment scores.

Figure S5. Correlation between PSR amino acid enrichment scores and codon GC content. Codon GC content was calculated as the average GC content of all codons for a particular amino acid. Plot shows the linear regression line with 95% CI. See Figure S7 for the PSR enrichment scores.

Figure S7. Occurrences of amino acids in the human sequence at the position of a PSR (positively selected site in the alignment). The y-axis represents enrichment scores comparing the PSR amino acid distribution (absolute numbers indicated) to a background distribution of amino acid occurring in all human sequences (11,096) used for our evolutionary analyses: $\text{ES}_{AA} = \log\left( \frac{\text{fraction}(AA_{psr})}{\text{fraction}(AA_{background})} \right)$.

Amino acid colors: blue = basic; red = acidic; green = polar; purple = neutral; gray = hydrophobic.
Figure S6. Distribution of PSG across the human genome and chromosomes. Individual PSG are marked by the red triangles to the left of the chromosomes; bars to the right indicate total numbers of PSG across the chromosomes. The numbers of PSG per chromosome are not significantly different from the expected numbers based on the total number of protein-coding genes per chromosome ($P \approx 0.1$, chi-squared test). Not in the figure: the PSG are distributed roughly equally across the different DNA strands (169/331 [51%] are on the minus strand, 162/331 [49%] are on the plus).
### Supplementary Tables

#### Table S1. Overview of genomes used in the analysis, taken from Ensembl release 78 (December 2014)

<table>
<thead>
<tr>
<th>Common name</th>
<th>Scientific name</th>
<th>NCBI Taxonomy ID</th>
<th>First genome (publication) available [ref]</th>
<th>Picture a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>Homo sapiens</td>
<td>9606</td>
<td>2001 [36,37]</td>
<td></td>
</tr>
<tr>
<td>Chimpanzee</td>
<td>Pan troglodytes verus</td>
<td>9598</td>
<td>2005 [333]</td>
<td></td>
</tr>
<tr>
<td>Gorilla</td>
<td>Gorilla gorilla gorilla</td>
<td>9595</td>
<td>2012 [398]</td>
<td></td>
</tr>
<tr>
<td>Orangutan</td>
<td>Pongo abelii</td>
<td>9601</td>
<td>2011 [399]</td>
<td></td>
</tr>
<tr>
<td>Gibbon</td>
<td>Nomascus leucogenys</td>
<td>61853</td>
<td>2014 [400]</td>
<td></td>
</tr>
<tr>
<td>Macaque</td>
<td>Macaca mulatta</td>
<td>9544</td>
<td>2007 [64]</td>
<td></td>
</tr>
<tr>
<td>Baboon</td>
<td>Papio anubis</td>
<td>9555</td>
<td>2012 b</td>
<td></td>
</tr>
<tr>
<td>Vervet (African Green Monkey)</td>
<td>Chlorocebus sabaeus</td>
<td>60711</td>
<td>2014 c</td>
<td></td>
</tr>
<tr>
<td>Marmoset</td>
<td>Callithrix jacchus</td>
<td>9483</td>
<td>2014 [401]</td>
<td></td>
</tr>
</tbody>
</table>

a Species pictures were taken from the UCSC Genome Browser [116], Ensembl [27] and https://commons.wikimedia.org/wiki/File:Charles_Darwin_photograph_by_Herbert_Rose_Barnaud_1881.jpg
b https://www.hgsc.bcm.edu/non-human-primates/baboon-genome-project
c http://www.genomequebec.mcgill.ca/compgen/vervet_research/genomics_genetics/
Additional Supplementary Tables are available at
http://www.cmbi.umcn.nl/~rvdlee/thesis/

- **Table S2.** 11,170 clusters of one-to-one orthologous genes in the nine primates.
- **Table S3.** 416 apparent Positive Selected Genes (aPSG) systematically inspected for artefacts.
- **Table S4.** 331 PSG: human genes with extensive statistical evidence for positive selection across nine primates.
- **Table S5.** 934 PSR: positively selected residues (codons) detected in the 331 PSG.
- **Table S6.** Function enrichment analysis of PSG using Babelomics (402).
- **Table S7.** Function enrichment analysis of PSG using DAVID (272).
- **Table S8.** Overlaps of the PSG with various virus-host interaction, immunity, and other datasets.
- **Table S9.** Known (antiviral) immunity genes detected to be under positive selection in small-scale evolutionary studies.
Transcriptome analysis of complex I-deficient patients reveals distinct expression programs for subunits and assembly factors of the oxidative phosphorylation system

*BMC Genomics (2015), 16 (1): 691

Robin van der Lee¹,*
Radek Szklarczyk¹,²,*
Jan Smeitink²
Hubert J.M. Smeets⁴
Martijn A. Huynen¹
Rutger Vogel³

¹Centre for Molecular and Biomolecular Informatics, ²Nijmegen Center for Mitochondrial Disorders – Radboud university medical center, Nijmegen, The Netherlands
³Department of Clinical Genetics, ⁴Unit Clinical Genomics – Maastricht University Medical Centre, Maastricht, The Netherlands

*Co-first authors
**ABSTRACT**

**Background:** Transcriptional control of mitochondrial metabolism is essential for cellular function. A better understanding of this process will aid the elucidation of mitochondrial disorders, in particular of the many genetically unsolved cases of oxidative phosphorylation (OXPHOS) deficiency. Yet, to date only few studies have investigated nuclear gene regulation in the context of OXPHOS deficiency. In this study we performed RNA sequencing of two control and two complex I-deficient patient cell lines cultured in the presence of compounds that perturb mitochondrial metabolism: chloramphenicol, AICAR, or resveratrol. We combined this with a comprehensive analysis of mitochondrial and nuclear gene expression patterns, co-expression calculations and transcription factor binding sites.

**Results:** Our analyses show that subsets of mitochondrial OXPHOS genes respond opposingly to chloramphenicol and AICAR, whereas the response of nuclear OXPHOS genes is less consistent between cell lines and treatments. Across all samples nuclear OXPHOS genes have a significantly higher co-expression with each other than with other genes, including those encoding mitochondrial proteins. We found no evidence for complex-specific mRNA expression regulation: subunits of different OXPHOS complexes are similarly (co-)expressed and regulated by a common set of transcription factors. However, we did observe significant differences between the expression of nuclear genes for OXPHOS subunits versus assembly factors, suggesting divergent transcription programs. Furthermore, complex I co-expression calculations identified 684 genes with a likely role in OXPHOS biogenesis and function. Analysis of evolutionarily conserved transcription factor binding sites in the promoters of these genes revealed almost all known OXPHOS regulators (including GABP, NRF1/2, SP1, YY1, E-box factors) and a set of novel candidates (ELK1, KLF7, SP4, EHF, ZNF143, and TEL2).

**Conclusions:** OXPHOS genes share an expression program distinct from other genes encoding mitochondrial proteins, indicative of targeted nuclear regulation of a mitochondrial sub-process. Within the subset of OXPHOS genes we established a difference in expression between mitochondrial and nuclear genes, and between nuclear genes encoding subunits and assembly factors. Most transcription regulators of genes that co-express with complex I are well-established factors for OXPHOS biogenesis. For the remaining six factors we here suggest for the first time a link with transcription regulation in OXPHOS deficiency.

**BACKGROUND**

Mitochondria are the primary source of cellular ATP, which is generated via electron transfer in the oxidative phosphorylation (OXPHOS) system using substrates derived from oxidation of carbohydrates, fatty acids and amino acids. The value of healthy mitochondria becomes evident in cases of OXPHOS deficiency. These metabolic disorders primarily affect tissues with a high ATP demand such as the brain, heart, and skeletal muscle, typically resulting in progressive energy deficiencies and childhood death [403]. Respiratory chain disorders occur in approximately 1:5,000-10,000 living births [404]. The most frequently encountered one is complex I deficiency (OMIM: 252010). No cure for OXPHOS deficiencies exists, and current interventions are either cumbersome
or only effective for specific types of the disease [405]. Furthermore, 40-70% of cases remain genetically unexplained [406,407] as no mutations are found in the genes encoding structural subunits or assembly factors, impeding genetic counseling. Therefore there is a great need for a better understanding of how the biogenesis and activity of the OXPHOS system is controlled.

Cells can control metabolic output by regulating gene expression. The OXPHOS system is constructed from a combination of nuclear and mitochondrial gene products, e.g. seven genes of complex I are encoded by the mitochondrial DNA and 37 by the nuclear DNA. This bigenomic assembly implies that there are at least two mechanisms for regulating OXPHOS gene expression: mitochondrial and nuclear. Replication, maintenance, and transcription of mitochondrial DNA are tightly regulated processes. Disturbances in any of these processes have been firmly linked to combined OXPHOS deficiency (for a recent review see [408]). In contrast, although much has been published about the relevance of metabolic (co)-regulators such as PGC-1α, NRF1, NRF2, YY1, and SP1 for the regulation of OXPHOS gene expression, little is known about the possible relationship between disturbed nuclear gene regulation and OXPHOS deficiency.

In this study, we investigate mitochondrial and nuclear gene expression patterns in patients with complex I deficiency under various conditions of perturbed mitochondrial metabolism. Gene expression clustering, co-expression calculations and analysis of transcription factor binding sites provide insights into nuclear transcription regulation of OXPHOS, suggesting regulation of the system as a whole rather than regulation of specific complexes. Our data also reveal that assembly factors follow an expression pattern that is more like genes encoding other mitochondrial proteins than like OXPHOS subunits. Finally, analysis of enriched regulators of nuclear genes co-expressing with complex I not only retrieves virtually all transcription factors (TFs) with a well-known role in the regulation of OXPHOS gene expression, but also identifies several factors not previously implicated in the regulation of OXPHOS in general or in respiratory chain disease.

**RESULTS**

**Culturing, incubation and RNA sequencing of complex I-deficient patient cells**

To investigate patterns of transcription in OXPHOS deficiency, we measured gene expression in two healthy fibroblast cell lines and two fibroblast patient cell lines carrying mutations in complex I genes NDUFS2 and ND5 (Figure 1). Cells were treated with vehicle (DMSO) and three compounds that stimulate or inhibit mitochondrial metabolism in order to trigger a transcriptional response of mitochondrial genes. Chloramphenicol inhibits mitochondrial gene translation and OXPHOS assembly and function [409]. Resveratrol stimulates mitochondrial growth/metabolism via SIRT1/AMPK [410]. AICAR stimulates mitochondrial metabolism as an AMPK agonist [411]. The compound incubations were done in duplicate, resulting in a total of 4 cell lines x 4 compounds x 2 replicates = 32 samples. Processing of the RNA sequencing data revealed expression values of 13,684 nuclear DNA encoded genes (referred to as nuclear genes) and 16 mitochondrial DNA encoded genes (referred to as mitochondrial genes) in all samples. We subsequently analyzed the transcriptomes using three approaches: gene expression clustering, co-expression calculations, and transcription factor binding site analysis (Figure 1).
Of our list of currently known 127 unique OXPHOS subunits and assembly factors (Table S1), all 13 mitochondrial and 112 nuclear OXPHOS genes were detected in the expression analysis. To investigate the distribution of OXPHOS genes in the 13,700 expression profiles we performed hierarchical clustering (Pearson uncentered, average linkage). Many OXPHOS genes share a distinct expression profile across the experiments and are significantly enriched among a sub-cluster of 1,518 genes (48/1,518 were OXPHOS, Fisher’s exact $P < 0.05$, Figure S1 and Table S2). Next we analyzed mitochondrial and nuclear OXPHOS gene expression separately.

**Differences in expression of mitochondrial OXPHOS genes between cell lines and treatments**

mRNA measurements for mtDNA-encoded genes are strongly correlated between biological replicates ($R^2=0.88$, see Methods). Total mitochondrial mRNA expression was not significantly different between the four cell lines (one-way ANOVA $P=0.6$) indicating that differences in the genetic background of our cells do not have a systematic influence on total mitochondrial gene expression. mRNA expression of mitochondrial genes in control vs. patient cells differed less than 1% implying that the complex I defects do not influence global mtDNA expression levels (Figure S2A). We did, however, observe a significant transcriptional response in cells upon perturbation of metabolism using different compounds (4 treatments, one-way ANOVA $P=0.0002$). The total mitochondrial mRNA expression was 15% higher upon chloramphenicol treatment compared to control DMSO, and significantly higher than any of the other treatments ($P<0.0003$ in pairwise comparison with the other groups, two-tailed paired T-test) (Figure S3). We observed high transcript abundance upon chloramphenicol treatments in all cell types. The variance in chloramphenicol-induced gene expression is very low compared to other treatments and the control condition, suggesting saturation of mitochondrial transcript abundance (Figure S3). Total mtDNA gene expression levels in AICAR and resveratrol treatments are lower than control 4% ($P=0.1$) and 5% ($P=0.001$), respectively.

mRNA levels of mitochondrial OXPHOS genes change in response to treatments (Figure 2A). These changes in transcript levels can be the result of a combination of factors, such as changes in mitochondrial transcription, in mitochondrial mRNA degradation rates, and in trafficking. The direction of mRNA changes upon treatments vary considerably between genes even...
Figure 2. Mitochondrial OXPHOS genes respond differentially to treatments and assembly factors tend to express differently from nuclear genes encoding OXPHOS subunits. Expression profiles of OXPHOS genes are shown in heatmap representation in 32 RNA sequencing measurements of control and complex I-deficient patient cells. Panel (A) shows the mitochondrial-encoded OXPHOS genes. Panel (B) shows the nuclear-encoded OXPHOS genes. At the top of the figure, controls, patients, and compound incubations are indicated across the samples, where numbers 1 and 2 refer to the cell line. On the right assembly factors and subunits per complex are labeled by color. On the left the central cluster of OXPHOS subunit genes is indicated by a vertical bar. Genes were clustered using average linkage clustering with uncentered Pearson correlation as distance matrix. On the bottom the horizontal bar depicts expression values. A value of 1.0 (black) denotes median log-expression of the gene (see Methods), with green denoting higher and red denoting lower expression levels.
within the same complex. For example, the transcript level of MT-ND2 increases by 50% upon chloramphenicol treatment compared to vehicle, while other subunits of complex I (MT-ND4, MT-ND5, MT-ND6) decrease up to 30% (Figure 2A and S4). We did not observe correlations between changes in mitochondrial gene expression and respiratory chain complex, position on the mitochondrial genome or mitochondrial strand of origin (data not shown). However, certain trends are apparent and reproducible. Up-regulation of genes encoding complex I subunits: MT-ND4, 4L, 5 and 6 upon AICAR treatment (~40%) and their down regulation in chloramphenicol (~10%) is significant compared to cells treated with control DMSO (P<0.05 or less for these genes, two-tailed paired T-test). MT-ND2 and, to a lesser extent MT-ND3 and MT-ATP8 show the opposite expression pattern: they are up regulated in chloramphenicol compared to AICAR (140-270%, P<0.05). Thus, mitochondrial gene expression is similar between cell lines but different upon treatment versus control vehicle, as subsets of mitochondrial OXPHOS genes respond differently to chloramphenicol and AICAR.

Differences in expression of nuclear OXPHOS genes between cell lines and treatments
Examination of the expression data for nuclear genes suggests differences in steady state and treatment-induced OXPHOS mRNA levels between control and patient cell lines, with control 2 generally showing higher nuclear OXPHOS gene expression than the other cell lines (Figure S2B). Furthermore, the chloramphenicol-induced accumulation of transcripts for a subset of mitochondrial OXPHOS genes (Figure 2A) is not matched by nuclear OXPHOS mRNA levels (Figure 2B). For nuclear complex I genes, unsupervised clustering captures similarity of transcriptional responses to AICAR and resveratrol in closely-knit clusters (Figure S6), but these responses are not consistent for all genes. Nevertheless, combining the gene expression profiles across all cell lines and treatment conditions is highly informative, as will be illustrated in the following sections.

Clustering identifies distinct expression patterns of OXPHOS subunits and assembly factors
Detailed analysis of the expression levels of nuclear OXPHOS genes revealed no specific clusters for individual OXPHOS complexes, arguing against complex-specific regulation of expression for the evaluated cell types and conditions (Figure 2B). However, clustering did highlight a distinction between the expression profiles of OXPHOS subunits and assembly factor genes: 67 of the 81 genes (83%) in the central largest cluster are subunit genes, while 28 of the 32 genes (82%) in the remaining smaller clusters are assembly genes (Figure 2B). Thus, although OXPHOS genes share similar expression profiles, subunits cluster together in a distinct group from OXPHOS assembly genes, indicative of differential transcription regulation.

Co-expression confirms absence of OXPHOS complex-specific expression profiles and differential behavior of assembly factors
To identify additional candidates with expression profiles compatible with OXPHOS genes, we calculated co-expression with complex I for all measured nuclear genes by integrating the gene expression profiles (Figure 1). As expected, known OXPHOS genes as a group have significantly higher co-expression with complex I than do other genes in the genome (Figures 3A and 3B, P = 3.1 × 10^{-28}, one-tailed Mann-Whitney U test). However, OXPHOS genes are also more co-expressed with complex I than other genes encoding proteins localized to the mitochondria [53] (P = 2.5 × 10^{-15}). Sub-classification of OXPHOS genes into assembly factors and the structural subunits of the
five complexes (Table S1) revealed no complex-specific co-expression patterns, in agreement with
the trends observed in the clustering approach above: complex I co-expression distributions are
similar for individual OXPHOS complexes (Figure 3B, $P = 0.33$, Kruskal-Wallis test). In fact, expression
profiles of complexes III, IV and V tend to be more similar to the average profile of complex I (i.e.
have higher median complex I co-expression scores) than those of complex I genes themselves,
supporting the notion that mRNA expression of OXPHOS genes is jointly regulated. The possible
exception is complex II, subunits of which tend to show less co-expression with complex I genes
(though not statistically significantly different) than do subunits of complexes III, IV and V.

Interestingly, assembly factors (including those involved in assembly of complex I itself) have a
significantly lower co-expression with complex I than structural subunits ($P = 1.8 \times 10^{-8}$, Kruskal-
Wallis test, Figure 3B) and are more similar to other mitochondrial genes, confirming that their
expression profiles are significantly different from OXPHOS subunits (Figure 2B). Thus, the
complex I-like expression profile seems specific for OXPHOS subunits compared to assembly factors
or other genes encoding mitochondrial proteins, but does not distinguish individual complexes.
Genes that consistently co-express with complex I under different conditions in different cells are likely to be functionally related, for example by being involved in OXPHOS biogenesis or its regulation. To identify such potential genes we considered the top 5% of complex I co-expression values, corresponding to a co-expression score cutoff of 0.54 (close to a local peak in the frequency distribution of co-expression scores, Figures 3 and S5, Table S3). This cutoff captures a total of 684 nuclear genes, including 43% (16/37) of the complex I subunits, 42% (15/38) of the other OXPHOS subunits, three assembly factors (NDUFAF3, ACAD9, C19orf79) and 96 other known mitochondrial genes. Functional classification of these top-ranking genes using DAVID [272] reveals a program that fits the biogenesis and breakdown of OXPHOS proteins (Table S4). Gene groups with the highest enrichment scores (ES) other than OXPHOS (37 genes, ES 12, mostly complex I/III/IV/V genes) are translation (51 genes, ES 23, mostly (mito)ribosomal genes), ribonuclease processing (12 genes, ES 8) and quality control (14/15 genes, ES 8/6, mostly ubiquitin and proteasome-related genes). That plasticity in transcription of metabolic/mitochondrial genes is tightly coupled to proteolytic breakdown of the involved activators was recently highlighted by Catic et al. [412] and could explain the enrichment of the latter group. Two additional enriched gene groups are of potential interest. The first group (ES 21) contains PRELID1 (mitochondrial morphology and function) [413], CHCHD2 (regulator of cytochrome oxidase) [414], ISOC2 (tumor development) [415], and MRPL53 (a mitochondrial ribosomal protein also found in [416] in a putative ribonucleotide complex with LRPPRC and SLIRP). The second (ES 7) contains 7 genes including the complex I assembly factor ACAD9, ETFa and ETFb (involved in beta oxidation of fatty acids), and C1QBP (mitochondrial protein synthesis) [417].

**Nuclear OXPHOS genes as a group are regulated by a common set of known and novel candidate OXPHOS transcription factors**

Next, we analyzed potential transcriptional regulators of the 684 (top 5%) nuclear genes having the highest co-expression with complex I. We obtained conserved transcription factor binding sites (TFBS) for 218 unique TFs in the promoter regions of 16,298 human genes from a comparative evolutionary analysis of the genomes of 29 placental mammals [30]. This information is independent of experimental conditions and cell type and is therefore typically well suited for exploratory analysis aimed at prioritizing which TFs may regulate a biological system (see Discussion). The data include at least one conserved TFBS in 606 of the 684 genes (89%) with the highest complex I co-expression values (Figure 1).

Genes that co-express with complex I are significantly enriched for 12 TFs compared to the background of all human genes after correction for testing multiple TF (Benjamini-Hochberg-corrected Fisher’s exact $P < 0.05$, Tables 1 and S5): GABP, Elk1, Nrf-1, Sp1, Klf7, Sp4, Ehf, znf143, Myc, YY1, E-box, and Tel2. Ten additional TFs are significantly enriched when they are evaluated individually (i.e. without correcting for multiple testing; Fisher’s exact $P < 0.05$, Tables 1 and S5).

Virtually all known OXPHOS transcription regulators are present among these 22 enriched factors, including GABP, NRF1/2, SP1, YY1, E-box. For example, the nuclear respiratory factors 1 and 2 (NRF1 and NRF2) have ranks 3 and 20 (out of 218 TFs tested, $P = 1.0 \times 10^{-15}$ and 0.033, respectively). Furthermore, NRF1 ranks first in the analysis of known complex I genes, and is also strongly over-represented in complex II-V genes and OXPHOS assembly factors (Table 2). The most significantly
enriched TF for complex I co-expressing genes is GABP/NRF2 (35% of co-expressing genes, 2.1-fold enrichment compared to all genes, P = 1.8 × 10^{-28}). NRF2 belongs to the ETS family of transcription factors, which also contains ETS1 (rank 17, P = 0.013). Other examples of known OXPHOS regulators of complex I co-expressing genes are SP1 (rank 4, P = 1.2 × 10^{-12}), YY1 (rank 10, P = 6.3 × 10^{-9}), CREB (rank 21, P = 0.041), the E-box regulatory motif (rank 11, P = 1.3 × 10^{-3}) and E-box factor c-Myc (rank 9, P = 2.7 × 10^{-5}). PGC-1-related coactivators (PRC), such as PGC-1α and PGC-1β, are known to act through PPAR(a) (rank 14, P = 8.6 × 10^{-7}) and PPARγ (rank 16, P = 0.012). Aside from known regulators, our analysis also identified six TFs not previously linked to OXPHOS gene expression (see

Table 1. Transcription factor binding site enrichment in promoter regions of the top 5% complex I co-expressing nuclear genes

<table>
<thead>
<tr>
<th>Rank</th>
<th>Transcription factor (TF)</th>
<th>TF occurrences in genome</th>
<th>TF occurrences in top 5% complex I co-expressing genes</th>
<th>Odds ratio</th>
<th>Fisher’s exact P</th>
<th>Benjamini-Hochberg-corrected Fisher’s exact P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GABP</td>
<td>2802</td>
<td>215</td>
<td>2.064</td>
<td>1.80E-28</td>
<td>3.92E-26</td>
</tr>
<tr>
<td>2</td>
<td>Elk1</td>
<td>2051</td>
<td>173</td>
<td>2.269</td>
<td>8.56E-27</td>
<td>9.33E-25</td>
</tr>
<tr>
<td>3</td>
<td>Nrf-1</td>
<td>2099</td>
<td>149</td>
<td>1.909</td>
<td>1.02E-15</td>
<td>7.38E-14</td>
</tr>
<tr>
<td>4</td>
<td>Sp1</td>
<td>7372</td>
<td>359</td>
<td>1.310</td>
<td>1.25E-12</td>
<td>6.79E-11</td>
</tr>
<tr>
<td>5</td>
<td>Klf7</td>
<td>4593</td>
<td>246</td>
<td>1.440</td>
<td>1.47E-11</td>
<td>6.40E-10</td>
</tr>
<tr>
<td>6</td>
<td>Sp4</td>
<td>4452</td>
<td>227</td>
<td>1.371</td>
<td>1.92E-08</td>
<td>6.98E-07</td>
</tr>
<tr>
<td>7</td>
<td>Ehf</td>
<td>774</td>
<td>61</td>
<td>2.120</td>
<td>2.25E-08</td>
<td>6.99E-07</td>
</tr>
<tr>
<td>8</td>
<td>znf143</td>
<td>630</td>
<td>50</td>
<td>2.134</td>
<td>3.63E-07</td>
<td>9.90E-06</td>
</tr>
<tr>
<td>9</td>
<td>Myc</td>
<td>1105</td>
<td>68</td>
<td>1.655</td>
<td>2.68E-05</td>
<td>0.00065</td>
</tr>
<tr>
<td>10</td>
<td>YY1</td>
<td>1366</td>
<td>74</td>
<td>1.457</td>
<td>0.00063</td>
<td>0.01378</td>
</tr>
<tr>
<td>11</td>
<td>Ebox</td>
<td>629</td>
<td>39</td>
<td>1.668</td>
<td>0.00127</td>
<td>0.02515</td>
</tr>
<tr>
<td>12</td>
<td>Tel2</td>
<td>1074</td>
<td>58</td>
<td>1.452</td>
<td>0.00264</td>
<td>0.04801</td>
</tr>
<tr>
<td>13</td>
<td>INSM1</td>
<td>503</td>
<td>31</td>
<td>1.658</td>
<td>0.00421</td>
<td>0.07067</td>
</tr>
<tr>
<td>14</td>
<td>PPAR</td>
<td>381</td>
<td>24</td>
<td>1.694</td>
<td>0.00855</td>
<td>0.13310</td>
</tr>
<tr>
<td>15</td>
<td>SF1</td>
<td>1173</td>
<td>59</td>
<td>1.353</td>
<td>0.01076</td>
<td>0.15631</td>
</tr>
<tr>
<td>16</td>
<td>PPARG</td>
<td>2093</td>
<td>97</td>
<td>1.246</td>
<td>0.01204</td>
<td>0.16208</td>
</tr>
<tr>
<td>17</td>
<td>Ets</td>
<td>2591</td>
<td>117</td>
<td>1.214</td>
<td>0.01264</td>
<td>0.16208</td>
</tr>
<tr>
<td>18</td>
<td>PU.1</td>
<td>839</td>
<td>44</td>
<td>1.410</td>
<td>0.01353</td>
<td>0.16383</td>
</tr>
<tr>
<td>19</td>
<td>KROX</td>
<td>2294</td>
<td>103</td>
<td>1.208</td>
<td>0.02223</td>
<td>0.25506</td>
</tr>
<tr>
<td>20</td>
<td>Nrf-2</td>
<td>1587</td>
<td>73</td>
<td>1.237</td>
<td>0.03256</td>
<td>0.35491</td>
</tr>
<tr>
<td>21</td>
<td>CREB</td>
<td>1000</td>
<td>48</td>
<td>1.291</td>
<td>0.04130</td>
<td>0.42878</td>
</tr>
<tr>
<td>22</td>
<td>HNF4</td>
<td>2119</td>
<td>93</td>
<td>1.180</td>
<td>0.04806</td>
<td>0.47620</td>
</tr>
</tbody>
</table>

Background: all human genes. Listed are all individually significantly enriched TFs (i.e. without correcting for multiple testing; Fisher’s exact P < 0.05). TFs with Benjamini-Hochberg-corrected Fisher’s exact P < 0.05 are shaded in grey. TFs with a known role in OXPHOS biogenesis are in bold. See Table S5 for detailed results.
As we have shown above, expression clustering and co-expression analysis indicate joint regulation of OXPHOS genes as a group, rather than specific regulation of individual complexes. We next asked whether there is also no evidence for complex-specific regulation in the conserved TF binding sites. Promoter regions of complex I subunits are significantly enriched only for Nrf-1 binding motifs (40.6% of complex I genes, 3.2-fold enrichment, Fisher’s exact $P = 8.3 \times 10^{-5}$, Table 2A). However, Nrf-1 is not specific for complex I genes as it is also over-represented in subunits of complexes II, III, IV and V (32.4% of complex II-V genes, 2.5-fold enrichment, Fisher’s exact $P = 0.030$, Table 2B) and in OXPHOS assembly factors (25.7% of assembly factors, 2.0-fold enrichment, Fisher’s exact $P = 0.02973$, Table 2C). Thus, it appears that there are no significantly over-represented transcription factors among the 218 tested that regulate specific OXPHOS complexes, confirming the patterns of general OXPHOS subunit regulation observed in the (co-)expression data.

**Discussion**

Disorders of the oxidative phosphorylation (OXPHOS) system are rare but devastating energy deficiencies. To date, the genetic basis of a large fraction (estimates range from 40-70%) of these disorders remains enigmatic [406,407]. When no mutations are found in any of the known OXPHOS

---

**Table 2. Transcription factor binding site enrichment in promoter regions of OXPHOS subunits and assembly factors**

<table>
<thead>
<tr>
<th>Transcription factor (TF)</th>
<th>A Complex I (n=32)</th>
<th>B Complexes II, III, IV and V (n=34)</th>
<th>C Assembly factors (n=35)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nrf-1</td>
<td>2099</td>
<td>2099</td>
<td>2099</td>
</tr>
<tr>
<td>zif143</td>
<td>630</td>
<td>1173</td>
<td>2802</td>
</tr>
<tr>
<td>Six</td>
<td>19</td>
<td>1078</td>
<td>562</td>
</tr>
<tr>
<td>SF1</td>
<td>1173</td>
<td></td>
<td>GABP</td>
</tr>
<tr>
<td>NR4A2</td>
<td>1078</td>
<td></td>
<td>2099</td>
</tr>
<tr>
<td>Nrf-1</td>
<td>2099</td>
<td>2099</td>
<td>2099</td>
</tr>
<tr>
<td>Elk1</td>
<td>2051</td>
<td></td>
<td>2802</td>
</tr>
<tr>
<td>Err-alpha</td>
<td>958</td>
<td></td>
<td>11</td>
</tr>
<tr>
<td>RXR</td>
<td>562</td>
<td></td>
<td>11</td>
</tr>
<tr>
<td>GABP</td>
<td>2802</td>
<td></td>
<td>11</td>
</tr>
<tr>
<td>Nrf-1</td>
<td>2099</td>
<td></td>
<td>2099</td>
</tr>
</tbody>
</table>

Background: all human genes. Listed are all individually significantly enriched TFs (i.e. without correcting for multiple testing; Fisher’s exact $P < 0.05$). TFs with Benjamini-Hochberg-corrected Fisher’s exact $P < 0.05$ are shaded in grey. Numbers in parentheses indicate the number of genes for which there is conserved TFBS data. Nrf-1 is marked in bolder font.
subunits and assembly factors, a possible explanation may be found in the genes that control their expression, such as transcriptional (co-)activators. A specification of which of these factors control OXPHOS gene expression, and how, would be helpful.

Large-scale gene expression analyses have previously revealed co-expression of genes involved in the OXPHOS system [87,89,90,418-421]. The associated transcription program is moderated by a set of (co-)activators, including PGC-1α, NRF1/2, YY1, and SP1. To our knowledge only one study focused on the possibility of individual expression programs for individual OXPHOS complexes [420]. In this study, only genes within OXPHOS complexes I and IV showed moderately higher degrees of co-expression with each other than with OXPHOS genes as a whole. However, no specific TFs (among 150 families tested), conserved across three organisms (human, mouse, rat), could be identified to explain this result.

We aimed to further investigate the existence of separate expression programs for individual OXPHOS complexes. To this end we investigated the transcription program for OXPHOS genes in complex I-deficient cells and assessed the regulatory elements involved. We used complex I deficient cell lines and controls in order to discriminate complex I-related expression responses upon drug treatment. To elicit a metabolic transcriptional response we incubated the cells with and without chloramphenicol, AICAR, and resveratrol. The effects of these incubations were generally similar between controls and patients. Chloramphenicol resulted in an accumulation of mitochondrial mRNA, likely due to the block in translation. Furthermore, after chloramphenicol and AICAR treatment, we observed contrasting changes in the expression of subsets of mitochondrial OXPHOS genes (COX1/ND4/ND4L/ND5/ND6 vs ATP8/ND2/ND3) (Figure 2A).

Chloramphenicol inhibits mitochondrial translation, hence the observed changes in transcript levels are likely the consequence of disrupted mitochondrial translation. Recent studies have highlighted feedback mechanisms between mitochondrial translation and transcription. For example, ribosome subunit MRPL12 interacts with mitochondrial polymerase POLRMT to enhance mitochondrial transcription [422,423]. In addition, POLRMT interacts with 12S rRNA methyltransferase h-mtTFB1 as a possible checkpoint for 28S and 55S ribosome assembly [424]. The changes in transcript levels that we observe upon inhibition of translation are not correlated with respiratory chain complex, position on the mitochondrial genome or mitochondrial strand of origin. How the abovementioned interactions could affect the levels of individual mitochondrial transcripts is unclear and likely partly controlled by regulatory proteins. A recent example of such a protein is FASTKD5, required for the maturation of a subset of mitochondrial OXPHOS mRNA’s, primarily COX1 [425].

Although our analysis of mitochondrial and nuclear OXPHOS gene transcription did not reveal complex-specific expression patterns, we did observe a significantly different expression profile for nuclear OXPHOS subunit genes versus other genes encoding proteins localized to the mitochondria, supporting differential nuclear gene regulation of a sub-mitochondrial process. Interestingly, OXPHOS assembly factors showed expression profiles that are significantly different from OXPHOS subunits: assembly factor expression tends to be more similar to non-OXPHOS mitochondrial genes. For example, the iron-sulfur cluster protein NUBPL (IND1) [426] is a complex I
assembly factor with an expression profile very different from OXPHOS subunits (Figure 2B) and low co-expression with complex I subunits (at ~0.3 is has the third lowest co-expression score of all 39 analyzed OXPHOS assembly factors). These findings suggest that expression of at least some assembly factors is controlled by other factors than that of subunits. This is perhaps not surprising considering that assembly factors can play multiple roles not exclusive to the biogenesis of OXPHOS complexes, for example in translation, membrane insertion, or the incorporation of prosthetic groups.

In our expression data of complex I-deficient cells, 684 genes represent the top 5% of nuclear genes that co-express with known complex I genes. Among these genes are many subunits of other OXPHOS complexes, confirming that different OXPHOS complexes have highly similar expression profiles. Other highly enriched gene groups in the top complex I co-expressing genes are those for translation and for quality control. Of particular interest are a number of genes implicated in RNA processing and a subset of fatty acid oxidation genes. For example, ACAD9 is essential for complex I assembly and plays no obvious role in fatty acid oxidation, despite a highly conserved fatty acid oxidation active site [427,428]. ACAD9 (Figure 2B) and two genes actually involved in fatty acid metabolism, ETFA and ETFB, have high co-expression with complex I subunit genes (scores ~0.6), while ACADVL, ACADM, and ACADS, which are evolutionarily related to ACAD9, all have lower scores of ~0.4. Co-expression of ACAD9 and a number of key fatty acid oxidation genes with complex I hints towards a possible functional link between these two metabolic pathways.

To explore which TFs may be important for regulating genes that co-express with complex I, we made use of a previously published data set of TF binding sites that are conserved across 29 mammals [30]. The conserved TFBS are detected solely on the basis of genome sequence and are therefore independent of experimental conditions and cell type [122]. In contrast, binding sites identified in for example ChIP-sequencing experiments, such as generated by the ENCODE consortium [29], are specific to cell type and experimental conditions. Although the conserved TFBS data has been shown to agree well with experimentally measured ChIP-seq binding sites [30], only a subset of TFs have been measured across many different cell types and conditions. Therefore, sequence-conserved TFBS data such as used in this study is typically well suited for prioritizing which TFs may regulate a biological system. Indeed, we also analyzed enrichment of TFs using TFBS data derived from ENCODE ChIP-seq peaks, either by creating various composite data sets that union all tissues and conditions measured, or by analysis of specific cell types relevant to mitochondrial functioning such as skeletal muscle and heart cells. The union data sets produced very large enrichments for almost all TFs tested, while the tissue-specific data lacked power and produced not a single over-represented TF. Thus, neither of these approaches, in our hands, were insightful for prioritizing TFs involved in complex I co-expression, or in fact for various other biological systems generally unrelated to mitochondrial function.

Analysis of conserved TF binding sites in promoter regions revealed 22 over-represented TFs compared with their genomic abundance. The enriched TFs (Table 1) correspond well with known OXPHOS regulators [429,430]. However, several over-represented TFs have not been previously implicated in the regulation of complex I or OXPHOS in general. For example, ELK1 is the second strongest enriched TF in genes that co-express with complex I ($P = 8.6 \times 10^{-22}$) and belongs to the
ETS family of transcription factors, which also includes known OXPHOS regulators NRF2, GABP, and ETS1. Interestingly, ELK1 has been linked to primary respiratory chain disease: its target genes show large differential expression between muscle cells and fibroblasts of patients [431]. EHF (ESE3, rank 7, \( P = 2.2 \times 10^{-6} \)) and TEL2 (ETV7, rank 12, \( P = 2.6 \times 10^{-1} \)) are two other ETS family members.

SP4 (rank 6, \( P = 1.9 \times 10^{-6} \)), together with SP1 and KLF7 (rank 5, \( P = 1.5 \times 10^{-11} \)) part of the Krüppel-like family of TFs, was recently implicated in the regulation of cytochrome c oxidase (OXPHOS complex IV) gene expression in primary neurons [432]. In addition, SP4 regulates the three mitochondrial transcription factors TFAM, TFB1M, and TFB2M, and the complex IV assembly protein SURF1. Thus, the high rank of this TF fits with its proposed role in OXPHOS gene regulation.

Zinc finger protein 143 (ZNF143; complex I co-expression rank 8, \( P = 3.6 \times 10^{-7} \); known complex I genes rank 2, \( P = 7.2 \times 10^{-3} \)) is a transcriptional activator for selenocysteine tRNA (tRNAsec). During mitochondrial respiratory chain dysfunction, ZNF143 upregulates tRNAsec, which results in increased expression of glutathione peroxidase 1 (GPX1) [433]. This mechanism has been proposed to protect cells from oxidative stress damage in conditions of respiratory chain dysfunction. In addition, ZNF143 binds to HCFC1, which is a common component of active CpG island promoters and coincides with YY1 and GABP, both relevant to OXPHOS biogenesis [434]. Taken together, ZNF143 is a strongly enriched regulator of genes that co-express with complex I across expression data of complex I-deficient patients. Further investigation of the transcription factors newly identified by our analyses may provide new clues towards gene regulation in deficiencies of the oxidative phosphorylation system.

**Conclusions**

To find new leads for explaining the many genetically unexplained cases of OXPHOS deficiency we have explored mitochondrial and nuclear gene expression and transcriptional elements of OXPHOS subunits and assembly factors in human complex I-deficient cells. We found that genes of the OXPHOS system co-express distinctly from other genes encoding mitochondrial proteins but found no support for distinct expression profiles for individual complexes. Genes encoding OXPHOS assembly factors follow an expression program different from that of OXPHOS subunits, suggesting that regulation of biogenesis occurs via different transcriptional activators. Many regulators of genes that co-express with complex I are well-established factors for OXPHOS biogenesis. However, for six factors, we suggest for the first time a link with transcriptional regulation of OXPHOS genes. The physiological relevance of these factors will need to be tested.
**METHODS**

**Cell culture and RNA isolation**
Control (internal culture no #4996 and #MW35) and patient (internal culture no #5170 and #9170) fibroblasts were cultured in M199 medium (Life Technologies) supplemented with 10% fetal calf serum (v/v) and penicillin/streptomycin. These cultures were exposed for 72 hours in the same medium with the addition of 125 μM chloramphenicol, 75 μM resveratrol, 500 μM AICAR, or 0.01% DMSO (vehicle). Duplicate treatments were harvested as two confluent T175 flasks (~20 million cells) for two control and two complex I-deficient patient cell lines, resulting in a total of 32 samples for RNA isolation.

**RNA sequencing**
RNA was isolated using the Purelink RNA Mini Kit (Life Technologies). RNA was treated with DNase and quality control was performed (OD260/280=1.8–2.2; OD260/230≥2.0; RIN≥7.0; 28S: 18S>1.0). Two 2 μg of RNA was analyzed using Illumina HiSeq2000 at least 10M clean reads per sample (BGI Genomics, Hong Kong).

**RNA-seq data analysis**
Raw sequence data was filtered for reads that contain adaptors in their sequence (0.5%) and with more than 50% of low quality base calls (quality value <= 5, 1.1% of total reads). 98.4% or 10,270,698 of short reads passed these criteria. After filtering the reads were mapped to the reference using SOAP2 pipeline ([435](#)) allowing for 1 or 2 mismatched bases. Between 80 and 89% of reads, depending on the measurement, were successfully mapped to the reference mRNAs, with 44%-52% uniquely to a single gene position. 18-24% gene transcripts were covered across their full length (90-100% of the transcript length) and 17-18% of genes were covered for less than 10% of their transcript length. Reads were subsequently mapped onto reference gene sequences to calculate RPKM (Reads Per Kb per Million mapped reads) ([436](#)). This procedure resulted in expression values in at least one of the 32 samples for 19,426 genes. Of these, 13,700 genes were present in multiple experiments that allowed further analyses. The raw RPKM values were log-transformed and median-centered (median=1.0). For analysis of the transcript levels of mtDNA-encoded genes we did not take into account non-protein coding genes, e.g., mitochondrially encoded 12S and 16S RNA (MT-RNR1, MT-RNR2) and tRNA leucine 1 (MT-TL1). Reproducibility of total mRNA measurements of the mtDNA encoded genes for the biological replicates was very high (R²=0.88, average difference between measurements 1.9%; 0.1%-12%). The transcript levels of mitochondrial genes were normalized in the same way as nuclear genes.

**Co-expression calculations**
For the complex I co-expression analysis we take into account all nuclear encoded subunits of the complex and its assembly factors NDUFAF1-4, TMEM126B ([136](#)), and ACAD9 ([427](#)). Pearson correlations of expression were calculated for each of the 13,684 genes with every gene that encodes a complex I subunit or assembly factor. The correlation values are based on the expression
measurements of the 32 samples. We then calculated for each gene the average co-expression with the extended complex I gene set (i.e., the genes mention above).

\[ c(g,S = \{NDUFA1, NDUFA2, \ldots\}) = \text{avg}_{g \in S} \rho(e[g], e[g']) \]

where \( c(g,S) \) is the co-expression of gene \( g \) with gene set \( S \), where \( S \) is the extended set of complex I genes and assembly factors, \( \rho \) denotes the Pearson correlation, and \( e[g] \) is the expression vector in 32 conditions and cell types for gene \( g \).

Similarly to expression patterns observed for all OXPHOS genes, complex I subunits and their assembly factors exhibit slightly divergent transcriptional programs. Figure S6 reveals that complex I core subunits cluster together separately from assembly factors. The first (and largest) group is composed mostly of core subunits (from NDUFA11, top, to NDUFA6). Two smaller categories, with respectively seven and six subunits, additionally contain complex I assembly factors (NDUFAF1, NDUFAF5 in the second subclass; TMEM126B, NDUFAF2, NDUFAF4 in the third). Resveratrol and AICAR experiments appear to be most discriminative for the subgroup classification as the first subclass (subunits) shows low expression in resveratrol and AICAR in patients, and higher expression in AICAR-treated controls (Figure S6). The smaller two subclasses of complex I genes do not exhibit such clear expression patterns. To account for the fact that subclasses fall under different transcriptional programs, we only considered Pearson correlations greater than 0. This increases the sensitivity of finding genes that fall under the transcriptional program of one of the subclasses (positive correlation), but not the other (negative correlation, which is not taken into account).

We selected a subset of 684 genes (representing the top 5% of complex I co-expression scores) for downstream analyses. We determined this cutoff by three criteria: (i) it captures almost half (42%) of the known OXPHOS subunits, (ii) it is close to a local peak in the frequency distribution of co-expression scores (Figures 3 and S5), and (iii) the size of the resulting set of genes (684) is well suited for enrichment analysis, both on the level of functional classification and gene ontology as well as in terms of conserved transcription factor binding sites (see below).

**Conserved transcription factor binding site data**

Data on conserved transcription factor binding sites (TFBS) in human were obtained from a comparative analysis of 29 genomes of placental mammals, such as primates, rodents and many farm animals [30]. In this study, transcription factor (TF) regulatory motifs were collected from the TRANSFAC [437] and Jaspar [113] databases, and several protein binding microarrays [105,438,439]. The presence of individual motif instances (putative TFBS) was predicted across the human genome based on conservation across the 29 mammals: for each motif match in human, the smallest phylogenetic subtree was calculated that contains the human motif and aligned motifs in other species [122]. TFBS were identified at a false discovery rate of 60% and show reasonable agreement with experimentally measured ChIP-seq binding sites [30].

To identify putative regulators of a gene, we used conserved TFBS in promoter regions of genes, which were defined as 4 kilobase (kb) windows centered at all annotated transcription start sites.
of the gene (i.e. 2kb upstream and 2kb downstream of each transcription start site). This approach identifies instances for 361 regulatory motifs corresponding to 218 unique TFs (most TFs have multiple similar binding motifs) in the promoters of 16,298 genes, with a median of 7 (average 9.19) unique TFs per target gene.

**TFBS enrichment analysis**

We used the Fisher’s exact test to calculate statistical over-representation of TFs regulating a gene set of interest compared to a background (e.g. all genes in the human genome, or all OXPHOS genes). Enrichment \( P \) values were corrected for testing multiple TFs (218 in total) using the Bonferroni or Benjamini-Hochberg false discovery procedures. TFs were judged to be significantly enriched at a significance level of 5%.

**Ethics**

The study has been carried out in the Netherlands in accordance with the applicable rules concerning the review of Commissie Mensgebonden Onderzoek Regio Arnhem-Nijmegen. The board has approved this study and patients have provided written informed consent.

**Availability of supporting data**


**ACKNOWLEDGEMENTS**

We are grateful to Pouya Kheradpour for providing conserved TF binding sites in human promoter regions and to Sergio Guerrero for assistance in hierarchical clustering. RvdL and MAH were supported by the Virgo consortium, funded by the Dutch government (FES0908) and the Netherlands Genomics Initiative (050-060-452). RS was supported by the Metakids Foundation. RV was supported by NWO VENI grant no 863.10.018.

**Authors’ Contributions**

RV performed cell culture, treatments, RNA extraction and analysis, and drafted the manuscript. RvdL analyzed co-expression data and transcription factor binding sites, and drafted the manuscript. RS carried out transcriptomics analysis and helped to draft the manuscript. HJMS contributed to the mitochondrial gene expression analysis. MAH helped draft the manuscript, and together with JS supervised the study. All authors read and approved the final manuscript.
Supplementary Figures

Figure S1. OXPHOS enriched cluster of genes. Heatmap representation of the expression of all detected genes in the 32 samples. One cluster of 1518 genes is significantly enriched for OXPHOS genes (Pearson correlation 0.995).

Figure S2. Mitochondrial OXPHOS genes respond differentially to treatments and assembly factors tend to express differently from nuclear genes encoding OXPHOS subunits (arranged by cell line). These are the data of figure 2 with samples arranged by cell line. In addition to the legend of figure 2: letters correspond to the treatment: D=DSMO, C=Chloramphenicol, A=AICAR, R=Resveratrol.
Figure S3. Total mitochondrial mRNA measured (RPKM) for mitochondrial protein coding genes. mRNA levels upon chloramphenicol (CAP) treatment are significantly higher compared to other conditions, denoted with a star (P-value 0.0003 or lower in pairwise comparisons, two-tailed paired T-test). Bars display RPKM measurements (mean ±SD) for all cell types and two biological replicates (total: 8 measurements).

Figure S4. mRNA levels of mitochondrial genes encoding complex I subunits. Each bar represents average RPKM measurements for all cell types and two biological replicates (total: 8 measurements).

Figure S5. Histogram (gray, left y-axis) and kernel density estimates (blue, right y-axis) of co-expression scores with known complex I genes. The red line marks a co-expression score of 0.54. The top 684 genes (from a total of 13,684) with higher scores than the 0.54 cutoff represent the top 5% of nuclear genes co-expressing with complex I. These genes are assessed in the functional classification and TF binding site enrichment analyses. Density estimates were calculated using a Gaussian kernel with a smoothing bandwidth of 0.005.
Supplementary Tables

Available at the BMC Genomics website (http://dx.doi.org/10.1186/s12864-015-1883-8) and at http://www.cmbi.umcn.nl/~rvdlee/thesis/.

- **Table S1.** Overview of the structural subunits and assembly factors of OXPHOS complexes as used in this study.
- **Table S2.** OXPHOS enriched cluster of genes.
- **Table S3.** Top five percent complex I co-expressing genes.
- **Table S4.** DAVID gene functional classification results of the top five percent complex I co-expressing genes.
- **Table S5.** TF binding site enrichment analysis of promoter regions of the top five percent complex I co-expressing genes (full results).

Figure S6. Dendrograms of the transcriptional response of genes encoding complex I subunits in 32 RNA-seq measurements. Complex I subunits and assembly factors NDUFAF1-4, TMEM126B and ACAD9 are included. Expression value 1.0 (black) denotes median log-expression of the gene (see Methods), with green denoting higher and red denoting lower expression levels.
Chapter 7

This chapter summarizes our contribution to, and is partly adapted from:

CiliaCarta: an integrated and validated compendium of ciliary genes

Submitted for publication

Teunis J. P. van Dam¹, Julie Kennedy², Robin van der Lee¹,
Erik de Vrieze³,⁴, Kirsten A. Wunderlich⁵, Suzanne Rix⁶, Gerard W. Dougherty⁷, Nils J.
Lambacher⁸, Chunmei Li⁹, Victor L. Jensen⁴, Michel R. Leroux⁶, Rim Hjeij⁷, Nicola Horn⁹,
Yves Texier⁸, Yasmin Wissinger⁹, Jeroen van Reeuwijk¹⁰, Gabrielle Wheway³,⁴, Barbara
Knapp⁵, Jan F. Scheel³, Brunella Franco¹², Dorus A. Mans¹⁰, Erwin van Wijk¹³, Françoise
Képès¹³, Gisela G. Slaats¹⁴, Grischa Toedt¹⁵, Hannie Kremer³,⁴,¹⁰, Heymut Omran⁷, Katarzyna
Szymanska¹¹, Konstantinos Koutroumpas¹³, Marius Ueffing⁹, Thanh-Minh T. Nguyen¹⁰,
Stef J.F. Letteboer¹⁰, Machteld M. Oud¹⁰, Sylvia E. C. van Beersum¹⁰, Miriam Schmidt¹⁰,¹⁶,
Philip L. Beales⁶, Qianhao Lu¹⁵, Rachel H. Giles¹⁴, Radek Szklarczyk¹, Robert B. Russell¹⁵,
Toby J. Gibson¹⁵, Colin A. Johnson¹¹, Oliver E. Blacque², Uwe Wolfrum⁵, Karsten Boldt⁶,
Ronald Roepman¹⁰, Victor Hernandez-Hernandez⁸, Martijn A. Huynen¹

¹Centre for Molecular and Biomolecular Informatics, ²Department of Otorhinolaryngology,
³Donders Centre for Cognitive Neurosciences, ⁴Department of Human Genetics
– Radboud university medical center, Nijmegen, The Netherlands
²Dublin, Ireland; ³London, ¹⁰Leeds, UK; ⁸Burnaby, BC, Canada; ⁵Naples, Italy; ¹³Evry Cedex, France
¹⁴Utrecht, NL; ⁹Mainz, ⁷Münster, ¹⁰Tübingen, ¹¹Heidelberg, ¹⁵Freiburg, Germany
Abstract

The cilium is an essential organelle of the human cell associated with a wide range of genetic diseases. The continuous discovery of new ciliary genes is improving our understanding of cilium function and pathomechanisms, yet a complete molecular parts list of cilia has not yet been achieved. As part of a large-scale effort to determine a compendium of ciliary genes, we generated and rigorously analyzed genomic, proteomic, transcriptomic and evolutionary data. We measured the predictive power of the data to identify ciliary localization and/or function, and systematically integrated seven datasets into a single probabilistic gene score using a Bayesian approach. This resulted in a set of 285 high-confidence new candidate ciliary genes. Our collaborators then identified ciliary associations for 24 out of 36 tested genes using six distinct experimental approaches, and characterized OSCP1 as a novel ciliary gene. Based on the Bayesian integration and experiments we estimate the total human ciliome to consist of about 1,200 genes. Our experimentally benchmarked set of candidates, combined with previous literature and annotation, forms the basis of the CiliaCarta, a comprehensive ciliary compendium of 836 genes with an estimated false discovery rate of 10%. The resource will be useful for objective selection of candidate ciliary genes for further study and for prioritization of disease-causing mutations in ciliopathy patients.

Background

Cilia are microtubule-based organelles extending from the surfaces of most eukaryotic cells, serving critical functions in cell and fluid motility, as well as the transduction of a plethora of sensory and biochemical signals [440]. Cilium disruption leads to a wide range of human disorders, or ciliopathies, characterized by defects in many different tissues and organs leading to symptoms such as cystic kidneys, blindness, bone malformation and nervous system defects [441]. Cilia are highly organized and compartmentalized structures, enveloped by an extension of the plasma membrane, with an internal and membrane composition that differs substantially from that of the rest of the cell [442]. So far, several hundred genes have been implicated in the formation and function of ciliary structures and associated signaling and transport pathways (600 genes in total between Gene Ontology [184] and the SYSCILIA Gold Standard [443]). It is likely that many more ciliary genes remain to be discovered, as new cases are reported on a regular basis, often associated to cilia-related genetic diseases.

Genomics datasets are rich sources of information to uncover the genes and proteins that constitute organelles and molecular systems (Chapters 1, 2, 6 and 8). Due to their complementary and potentially conflicting nature, probabilistic or weighted integration of independent data sources is especially powerful. Bayesian integrative genomics has successfully been applied to various cellular systems such as small RNA pathways [52] and the innate antiviral response (Chapter 2 [326]), and most notably in MitoCarta [53,189], a compendium of mitochondrial proteins that has been extensively used by the biomedical community. Like the mitochondrion, the cilium has been investigated for signals in individual genomics data to predict new ciliary genes, e.g. the specific occurrence of genes in species with a cilium [444], the presence of TF binding sites like the X-box...

150 | Chapter 7
motif [445], and the co-expression of ciliary genes [446]. This chapter describes our contributions to the study design, and the generation, analysis and integration of diverse genomics datasets as part a big effort to achieve a compendium of ciliary genes.

**Results**

**Generation and analysis of seven datasets predictive for the cilium**

We constructed five new datasets from proteomics, genomics, expression and evolutionary data and complemented these with two public datasets (Table 1). Datasets were selected to ensure independence of the types of evidence and comprehensive coverage of molecular signatures for ciliary genes. Each dataset contains a highly significant signal for discovering ciliary genes ($P$ values range from $10^{-14}$ to $10^{-69}$, Table 1).

**Protein interactions.** Based on three distinct protein-protein interaction (PPI) datasets, we inferred proteins to interact with ciliary components as a proxy for being part of the cilium. First, our collaborators performed mass spectrometry (MS) of ciliary protein complexes: one approach involved tandem-affinity purification (TAP-MS [447]; 181 ciliary proteins used as baits), the second used a more sensitive and quantitative method (SILAC-MS; 16 baits, 14 of which were also part of TAP-MS). Because of their similar methodology and large bait overlap, we merged both studies into a single dataset (Mass spec-based PPI), which identified 4,799 unique proteins interacting with 184 ciliary proteins. In a second independent approach, our collaborators measured binary PPIs in yeast two-hybrid (Y2H) screens of cDNA libraries derived from ciliated cell lines and tissues (brain, kidney, retina, testis). 343 proteins were identified to interact with 27 baits.

**Gene regulation.** The RFX and FOXJ1 TFs play an important role in ciliogenesis [448,449]. Their DNA binding motifs have been used to predict novel ciliary genes in nematode and fly [448,449].

**Table 1. Coverage and predictive power of the cilium datasets**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of genes in dataset</th>
<th>Coverage of genome</th>
<th>Coverage of Gold Standard</th>
<th>$P$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mass spec-based PPI</td>
<td>4799</td>
<td>21.1%</td>
<td>65.9%</td>
<td>$2.0 \times 10^{-64}$</td>
</tr>
<tr>
<td>TAP-MS</td>
<td>4410</td>
<td>19.4%</td>
<td>64.9%</td>
<td>$1.5 \times 10^{-67}$</td>
</tr>
<tr>
<td>SILAC-MS</td>
<td>1397</td>
<td>6.2%</td>
<td>21.5%</td>
<td>$3.8 \times 10^{-10}$</td>
</tr>
<tr>
<td>Y2H PPI</td>
<td>343</td>
<td>1.5%</td>
<td>9.2%</td>
<td>$2.1 \times 10^{-14}$</td>
</tr>
<tr>
<td>RFX/FOXJ1 TF binding sites</td>
<td>2201</td>
<td>9.7%</td>
<td>29.4%</td>
<td>$2.2 \times 10^{-22}$</td>
</tr>
<tr>
<td>Cilia co-occurrence (DDP ≤ 9)</td>
<td>1485</td>
<td>6.5%</td>
<td>30.5%</td>
<td>$2.8 \times 10^{-17}$</td>
</tr>
<tr>
<td>Expression screen (S ≥ 1.5)</td>
<td>5448</td>
<td>24.0%</td>
<td>75.5%</td>
<td>$2.4 \times 10^{-69}$</td>
</tr>
<tr>
<td>Liu et al. 2007 (cilium protein content, proteomics)</td>
<td>2085</td>
<td>9.2%</td>
<td>38.4%</td>
<td>$1.4 \times 10^{-43}$</td>
</tr>
<tr>
<td>Ross et al. 2007 (ciliogenesis expression dynamics)</td>
<td>1204</td>
<td>5.3%</td>
<td>26.2%</td>
<td>$2.4 \times 10^{-43}$</td>
</tr>
</tbody>
</table>

Most datasets have close to genome-wide coverage in terms of which genes/proteins could potentially have been measured by the approach. Coverage columns denote the fraction of the genome or SYSCILIA Gold Standard ciliary genes actually identified by each approach. $P$ values indicate significant overrepresentation (Fisher’s exact test) of the Gold Standard compared to the genomic background. Mass spec-based PPI represents the union of TAP-MS and SILAC-MS, resulting in seven datasets for integration.
processed the publicly-available data from 29 mammalian genomes [30] (Chapters 2 and 6) to obtain human genes with evolutionarily conserved RFX (X-box) or FOXJ1 TF binding sites (TFBS) in their promoters (4kb windows around transcription start sites). The requirement for conservation in other mammals infers a high level of confidence that the motifs are relevant and not spurious hits. We identified RFX and/or FOXJ1 motifs in 2,201 human genes, which are strongly enriched for a ciliary function compared to the rest of the genome (3-fold enrichment, \( P = 2.2 \times 10^{-22} \), Table 1).

**Co-expression.** In expression screening [90] separate gene expression datasets are weighted for their potential to predict new genes for a system by measuring the level of co-expression of the known genes. We have already successfully applied this approach to predict TMEM107 as part of the ciliary transition zone [51] and now extend the approach to the complete cilium. An integrated cilium co-expression dataset was constructed by applying the weighted WeGET method [91] to ciliary genes, making use of 465 human expression datasets [28].

**Co-evolution.** Given the number of independent losses of the cilium in eukaryotic evolution [450], gene presence/absence profiles have a high value for predicting new cilium genes [51,451]. We constructed a comprehensive gene-cilium co-evolution dataset by correlating the occurrence of orthologs of 22,000 human genes in 52 eukaryotic genomes to that of cilia or flagella using differential Dollo parsimony (DDP)[49]. Thus we obtained an objective measure for each human gene that describes how well its evolutionary trajectory (e.g. its origin and loss events) matches that of the ciliary system. More than half (54%) of the 48 genes that best match the evolutionary cilium profile (two to four mismatching events) are known ciliary genes.

**Public proteomics and expression data.** In addition to the in-house data and calculations we included two published high-throughput datasets. The first is a proteomics dataset from Liu et al. [452], who characterized the protein content of sensory cilia derived from murine photoreceptor cells. The second is a gene expression dataset from Ross et al. [86], who determined genes up-regulated during cilium formation (ciliogenesis) in human lung epithelial cells. Both datasets have a high coverage compared to other published data at similarly high predictive value (Table 1).

**Bayesian integration provides gene-specific probabilities for cilia involvement**

By combining the complementary sources of evidence for cilium function, we can in principle, obtain a data-driven, objective, high-confidence compendium of ciliary genes. To integrate the datasets in a probabilistic manner we assessed their efficacy at predicting ciliary genes using (i) the SYSCILIA Gold Standard, a manually curated set of 302 known ciliary genes (the ‘positive’ set)[443], and (ii) a set of genes whose proteins show non-ciliary subcellular localization and are most likely not involved in ciliary function (the ‘negative’ set; 1,275 genes with GO Cellular Component terms - extracellular, lysosome, endosome, peroxisome, ribosome, or nucleolus). We divided each dataset into appropriate sub-categories that reflect increasing likelihoods to report ciliary versus non-ciliary genes (Figure 1A). The contributions of all datasets were then summed up per gene, and the final log-summed likelihood ratio, which includes the prior expected probability to observe a ciliary gene in the genome, we here call the CiliaCarta score. This score therefore represents the likelihood for a gene to be ciliary based on all data considered. The integrated CiliaCarta score readily distinguishes between the positive and negative set (Figure 1B, \( P = 2.8 \times 10^{-85} \), Mann-
The top ranking genes are highly enriched for known ciliary genes (Figure 1C). The Bayesian classifier outperforms any individual dataset while achieving genome-wide coverage (Figure 1D, area under the ROC-curve = 0.86).

**Experiments confirm the high predictive value of the integrative CiliaCarta score and characterize OSCP1 as a novel ciliary gene**

In order to validate the quality of our predictions, our collaborators performed a battery of experimental tests. Our inclusion threshold (estimated false discovery rate [FDR] of 25%) resulted in 404 genes, 285 of which were not in the training sets and thus constitute novel candidate ciliary genes. Appraisal by the cilia experts in our team revealed that many of these genes actually have known ciliary functions.

36 genes were tested in six distinct approaches to investigate ciliary localization and function in human, mouse, zebrafish and *C. elegans* (nematode). Combining the results of all the validation experiments, our collaborators observed ciliary associations for 24 out of 36 genes tested. The
experimental positive predictive value (PPV) of the Bayesian classifier thus approaches 67% (Figure 2). This value exactly matches the theoretical PPV of 67% (FDR of 33%, excluding genes from the training sets, Methods) calculated for the 285 candidate genes using the integrated CiliaCarta score. The experimental validation rate is significantly higher than expected from the estimated prior probability of finding a ciliary gene in the human genome (Observed = 24 / 36 (67%), Expected = 1.8 / 36 (5%), \( P = 3.1 \times 10^{-23} \), hypergeometric test).

Our collaborators then performed deeper functional characterization of one promising novel ciliary gene, **OSCP1** (organic solute carrier partner 1, CiliaCarta rank 402), which had previously been implicated in various non-ciliary functions including inflammation, apoptosis and tumor suppression [453]. **OSCP1** was identified to be ciliary in four species and cause a cilium dysfunction phenotype in zebrafish. Thus, independent experiments have benchmarked our Bayesian classifier and indicate its high value for identifying novel ciliary genes.

**Discussion**

In this work, we have presented an integrated, systems-level analysis of the genes that make up the cilium organelle. Organelles are typically coherent molecular systems, physically separated from the rest of the cell, which makes them excellent targets for systematic description by high-throughput genomics data. By integrating seven heterogeneous large-scale datasets dedicated to describing the cilium we have added 228 probable cilia-related genes to the human cilium repertoire, expanding previous annotation by 38% to 836 human genes (Figure 3). We put forward

![Figure 2](image_url)
these ciliary genes as the “CiliaCarta” resource, an experimentally benchmarked compendium of genes involved in ciliary structure, function and development (available in Van Dam et al. [454]).

CiliaCarta is likely to be incomplete still. Current estimates for the total number of ciliary genes range from one to two thousand genes depending on techniques used and species studied [455]. Our Bayesian analysis allows us to obtain a first substantiated estimate, based the combined evidence at multiple molecular levels and techniques. Using the posterior probabilities and the outcome from the validation experiments we estimate the size of the ciliome to be approximately 1,200 genes (Methods).

Our predictions are relevant to human disease. KIAA0753, part of our high-confidence ciliary candidates, was recently shown to interact with the cilium and give rise to Oral-Facial-Digital syndrome ciliopathy [456]. In conclusion, the genome-wide CiliaCarta score should make it possible to efficiently and objectively prioritize candidate genes to facilitate the discovery of new cilium biology and to identify the genetic causes of cilia-related diseases.

Figure 3. The CiliaCarta resource (836 genes). The SYSCILIA Gold Standard (302) and Gene Ontology (510) currently cover 608 human genes. Our Bayesian integration adds 228 probable ciliary genes.
Methods

Construction and availability of individual genomics datasets and validation experiments are described in detail in Van Dam et al. [454]. Technical aspects, assumptions, and performance calculations of the naive Bayesian integration approach are described in detail in Chapter 2. We pursued independence of evidence by integrating datasets that are methodologically largely unrelated (though biologically correlated by virtue of their ability to identify ciliary genes). To prevent overestimation of predictive value through circular reasoning, we adapted the positive training set for the PPI datasets as well as the expression screen as follows:

- Many of the PPI baits are known ciliary components and thus part of the positive training set. We excluded these bait proteins from the positive set used for evaluating the Y2H, TAP-MS and SILAC-MS data.
- The expression screen was evaluated using ciliary components from GO (GO:Cilium), excluding genes that also occur in the positive set. We also excluded from the expression screen the dataset from Ross et al. [86], because it is already included in the Bayesian classifier as a separate dataset.

False discovery rate calculations

False discovery rates (FDR) were corrected to account for deviations in positive and negative gene set sizes compared to the actual populations of positives and negatives for the whole genome (Chapter 2). We considered the top 404 CiliaCarta ranked genes (FDR of 25%) for further studies. After excluding known (non-)ciliary genes from the training sets, this resulted in 285 high-confidence novel ciliary candidates. We obtained the FDR for the remaining 285 genes by: (i) calculating the expected true and false positives among the top 404 genes based on the 25% FDR threshold (TP = 75% of 404 = 303; FP = 25% of 402 = 101), (ii) subtracting training set genes (TP - 112 positive genes = 191; FP - 7 negative genes = 94), and (iii) using the adjusted TP and FP to recalculate the FDR = 94 / (191 + 94) = 94 / 285 = 33%.

Ciliome size estimation

Accurate calculation of posterior probabilities and false discovery rates requires an a priori expectation of how many ciliary genes there are in the human genome. We can reasonably assume this ciliome to be much bigger than the 608 human genes currently annotated as ciliary (Figure 3), but to our knowledge no substantiated estimate exists. We have chosen a prior that we deem to be both reasonable and conservative: 5% of all human genes (i.e. 1,135 of 22,693).

Our Bayesian framework, combined with the experimental validation rate, can be used to obtain a systematic posterior estimate for ciliome size in two ways. The first approach would make use of a discrepancy between the expected and observed FDRs. Remarkably, in our study these FDRs are approximately equal (i.e. 33%), which means that our prior ciliome size estimate (1,135) is accurate. The second approach takes the specific CiliaCarta scores per gene into account. We considered each ciliary candidate to be a random variable with a binary outcome (success, a true ciliary gene, or failure, not a ciliary gene) whose probability of success is defined by the posterior probability [CiliaCarta score, P(ciliary gene|data)]. Effectively this corresponds to a Bernoulli process with different probabilities for success or failure at each trial: assuming independence, the expected
value $E$ (successes) for a set of $n$ binary random variables (i.e., ciliary candidates) equals the sum of the expected values of the individual variables (i.e., its probability of success). This resulted in an expected 1,273 ciliary genes. Averaging the two estimates (1,135 and 1,273) we arrive at a ciliome size of approximately 1,200 genes. While these estimates rely on various factors such as the prior and posterior probabilities, an independence assumption, and the validation experiments, they are based on evidence from a variety of data systematically combined using probability theory, which makes them the most substantiated to date.

**Authors’ Contributions**

TJPvD and MAH conceived and led the study. TJPvdD and MAH, wrote the manuscript with significant contributions from RBR, OEB, RHG, and RvdL. RvdL and TJPvD developed the algorithm and performed the bioinformatic analyses. JK, EdV, KAW, SR, GWD, NJL, CL, VLJ, RH, and VH-H performed the validation experiments and experiments on OSCP1. EdV, NH, YT, YW, JR, GW, BK, JFS, DAM, EvW, GGS, KS, TMTN, SJFL, SECvB, and KB performed experiments leading to the Y2H and SILAC data sets. TJPvD, RvdL and RS created the bioinformatics data sets. TJPvD, JvR, KK, GT, QL collected, quality assessed, formatted and mapped the data sets. MRL, FK, HK, HO, MU, PLB, BF, MS, RHG, RBR, TJG, CAJ, OEB, UW, KB, RR, VH-H, GWD, and MAH suggested strategies and supervised work.
Chapter 8

General Discussion
Systematic definitions of molecular systems

Systematic organization of proteins and genes into associated groups (e.g. functionally, molecularly, evolutionarily, phenotypically-related) is useful for studying biology, but defining such systems is not always trivial. While some systems represent clear biological entities, others can be subjective, historical and vague, and should in fact be considered as modules within cellular molecular networks as a whole [25,181]. For instance, ‘distinct’ innate immune pathways are actually highly interconnected and engage in extensive crosstalk to integrate a multitude of molecular signals, such as various pathogen patterns, cytokines, and cell-cell contacts [14]. Furthermore, system behavior and composition are dynamic and depend on a plethora of factors including the organism, cell and environment [11,17,25,181]. At the level of individual system components, there is still insufficient awareness that genes can have multiple, widely different functions, potentially in so-far unrelated systems; that is, evidence for one function does not rule out another. Despite the simplifications, studying defined systems in isolation is valuable for understanding both the individual systems as well as their role in the bigger scheme of living things.

System coherence determines predictability

Systems differ intrinsically and widely in how well they can be characterized. Organelles such as the mitochondrion and the cilium are physically separated from the rest of the cell and represent clear biological entities that can be defined by their total protein content [53,443](Chapter 7). In contrast, definitions of intracellular signaling networks such as innate antiviral pathways [14](Chapter 2) or disease systems such as intellectual disability [457] tend to be more fuzzy because they are part of much larger interconnected molecular networks. In our experience, the more biologically coherent the components of a system are, the better the system can be captured and described by genomics data. In turn, this ‘describability’ relates directly to the ‘predictability’ of the system, such as the ability to identify components through integrative genomics approaches. Nevertheless our studies and other recent publications (e.g. [54,180,192], Chapter 1) suggest that we are reaching the capability to predict components for even the more complex, heterogeneous systems, given sufficient quantities of the right data.

Mechanistic function prediction remains challenging

Current bioinformatics technology thus allows the inference of associations between genes and systems with fairly high confidence. Once we suspect that a gene is involved in a certain biological process, however, deeper characterization of its precise role [184,458] requires disproportional effort, especially for mechanistic detail. Clues about where in the identified process a gene may function may be acquired from protein interactions, co-expression patterns and other sorts of information [32]. For example, after identifying numerous novel genes involved in antiviral signaling, for some we could propose where in the network they might function based on known PPIs (Chapter 2). Co-expression and co-evolution analyses could locate TMEM107 not only to the cilium but
more specifically to the ciliary transition zone [51]. On a mechanistic level, analysis of sequences and homology may suggest molecular function based on protein domains [209], peptide motifs [459] and disordered regions [239]. When available, structures or homology models of proteins and protein complexes may provide detailed molecular insights [460], for instance allowing the mapping of interaction interfaces (e.g. in combination with sequence co-variation approaches [461]) and positions known to be important for functionality (e.g. mapping positively selected positions in virus-receptor interactions, Chapter 5). None of these approaches are guaranteed, however, to provide (reliable) predictions about detailed functionality, meaning that experimental characterization is often the only way to discover where and how a gene functions in a process. As mechanistic whole-cell models are a distant prospect [462], we will remain dependent on such detailed studies for deciphering the functions of single genes.

Understanding big data to progress biology
Rapid developments in techniques and availability of heterogeneous, large-scale molecular data have been the major drivers enabling the research presented in this thesis. Many of the measurement techniques and certainly the quantities of data we have used, have been developed and obtained only during the last decade or so. The scientific community increasingly recognizes the potential of analyzing and combining these big data for patterns relevant to any biological system of interest [463-467]. Biology is turning into a data-driven science, similar to the earlier developments in, for example, astrophysics and climate science [19-21]. We believe that, like in the last decade, future progress in the description of complex molecular systems will come mainly from better and more data.

Recycling and recombining the deluge of biological data
Thorough processing and interpretation of data in order to generate knowledge typically requires more time and effort than generating the data in the first place (e.g. [21,468,469]). It is therefore likely that not all knowledge has been fully extracted from most existing datasets. In principle, computational analyses of available data are complementary to experimental studies that keep generating novel data. However, high-throughput experiments remain expensive [465] despite technological advancements and the plummeting the costs of sequencing in particular [469]. Significant costs and challenges are not just associated with data generation but also with data processing steps, including quality assessment, storage and computation [21,465,469]. Data reanalysis is one way to maximize these investments as it capitalizes on such previous expenses related to data generation and processing. Thus, exploiting available data is an attractive option not only because it has the potential to give a more complete picture of biological systems than do individual experimental approaches, but because it is comparatively cost-effective as well.

New insights can be gained by studying available data from a perspective other than the questions for which they were originally measured, especially when data are combined in large collections (Table 2 in Chapter 1). In Chapter 2, for instance, we made use of 10,000 gene expression datasets [28,470], networks of 75,000 human-human [31] and 30,000 human-virus protein interactions [33], comparative analysis of 33 genomes [27,63,209], and genome-wide binding sites for 218 transcription factors [30]. Data availability is an obvious but vital requirement for re-use, but it is not uncommon for valuable datasets to remain hidden in the supplementary data of publications. Efforts that facilitate
public sharing and standardized deposition of data are important [197,471], as are frameworks that allow appropriate attribution to the original contributors [472]. Analogies between independent reanalysis of data and stealing are obviously not helpful and can only hamper such efforts [473,474]. Our studies reiterate the usefulness of data re-use, or ‘data recycling’, for discovering novel biology - existing data are not waste -, and contribute to the justification of biological big data.

**Knowledge-based exploration of data for better predictions**

All chapters of this thesis depend on a solid biological understanding of the system of interest. Constructing good datasets for predicting genes involved in a system is naturally most successful when guided by prior knowledge about the system and its expected behavior (e.g. [443,475,476]). Such datasets are often not readily available in the specific form that would suit the system under study and require customization based on what is known about the system. For instance, we used key known transcription factors, rather than all TFs for which data is available, for innate immune pathways, cilia, or mitochondria to find target genes likely also involved in these systems. Similarly, given previously observed viral antagonism of the antiviral immune response [70,201], we constructed a dataset of virus-interacting human proteins and examined genes that reduce infection rates. Besides the data themselves, curated gold standards of genes known to be involved in a system are a vital aspect determining the success of an integrative genomics approach. Involvement of experts from the very start has benefitted not only the design of these computational aspects of our studies, but has also stimulated early discussions about possible follow-up experiments. Importantly, these early collaborations allowed for an optimal fit between computations and experiments. For instance, the predictions for antiviral response regulators in Chapter 2 were motivated partly by the opportunity offered by our collaborators to develop an experimental system for testing viral RNA-induced production of IFNβ. Our cilium gene predictions in Chapter 7 were even tested by different research groups in multiple model organisms.

The other-end alternative to our knowledge-based approach would be to identify predictive datasets through automated assessment of large collections of available data (e.g. [10], Discussion in Chapter 4). A recent study applied such an approach to multiple immune pathways at the same time [192]. While this does still depend on previous knowledge of pathway members, a major difference with our Chapter 2 study of the RLR pathway is that the assessed data are generic and not tailored specifically to the pathway. In a continuum of integrative genomics approaches, we argue that fully automated approaches are more widely applicable while knowledge-based approaches have proven to deliver more specific predictions.

**Prediction prowess of genomics data**

The Bayesian integration approach weighs datasets, thereby statistically increasing the contribution of highly predictive datasets compared to less useful datasets. This means that any number of datasets of any predictive strength can theoretically be combined in a balanced fashion. Our studies have used heterogeneous types of genomics data. Unsurprisingly, we (Chapters 2 and 7) and others [32,52,189] found that some data types are generally better than others at inferring functional associations between genes. Protein interactions and co-expression patterns tend to be most informative. Other valuable data types include co-evolution, sharing of protein domains, and conservation of TF binding sites (e.g. CMEM method, Chapters 1, 2, 6 and 7). Datasets derived
from genome-wide siRNA screens tended to be only marginally predictive in our hands, including the antiviral host factors in Chapter 2 and a screen for ciliogenesis regulators related to Chapter 7 [477]. In contrast, medium-scale siRNA screening did validate half of our Chapter 2 predictions. The limited value of genome-wide siRNA datasets may partly be explained by the typically poor target-specificity of siRNAs (e.g. [478]) or by non-optimal application of these data. For instance, virus-host siRNA screens tend to focus on factors required by viruses and may have limited sensitivity in finding antiviral genes. In general, the differences in predictive power between data types may reflect (i) the inherent limits and maturity of the underlying measurement techniques, and (ii) the importance of the measured molecular level for determining cellular states [479], with localization data and protein interactions perhaps providing the most direct evidence for inferring functional associations.

Dedicated data, measured with the specific aim to characterize one particular system, understandably tend to be most powerful for studying that system. Examples include the mitochondrial and ciliary protein content from mass spectrometry, gene expression measurements under conditions known to induce the system of interest, and genetic interactions using a specific pathway readout. However, dedicated datasets are not always better than derived data, nor are dedicated data always available. Coverage is another factor influencing comprehensive description of a system: highly predictive data may cover a smaller fraction of genes than data with lower prediction power. While the trends outlined here are based on experience from multiple studies, their generalization should be subject to caution as the performance of a data type obviously depends on the individual dataset and the system under study. All in all, data integration has the potential to amplify the inherent strengths of individual data, while diluting their weaknesses.

**Rapid developments in large-scale molecular characterization**

Ongoing technological developments will further improve the characterization of cells and organisms, thereby offering further opportunities to comprehensively study molecular systems. Genome sequences will be obtained for more species, individuals and populations, using longer reads with higher quality [35,38]. This will improve the resolution of comparative analyses studying various aspect of genome evolution including the detection of conserved regulatory elements [74] (Chapter 1) and the inference of positive selection (Chapter 5). Measurements of gene expression, histone modifications and transcription factors in more cell types will improve the understanding and predictions of gene regulation [480]. Protein interaction studies will cover larger fractions of cellular proteomes in more conditions, and will profile intact protein complexes and post-translational modifications [77], increasingly at atomic detail [481]. CRISPR-based methods are causing a revolution by enabling rapid large-scale genotype-phenotype screens through genome editing [154]. Comprehensive measurements of genetic interactions would be just one application [482]. Single-cell methods will continue to replace bulk measurements and provide insights into cell-to-cell variability through high-resolution characterization of gene expression levels and epigenetic marks [97]. High-throughput imaging techniques will provide greater flexibility in measuring the cellular effects of perturbations and the localization patterns of proteins [483]. Following trends in, for instance, cancer (e.g. TCGA [484]) and immunology [3], more cellular and disease systems will undergo multi-level molecular characterization. These directions represent just some of the most
promising developments in large-scale cellular and molecular measurements. Together they will undoubtedly continue the technological revolution in biology.

**Proper practices for processing data and scrutinizing results**

Evidently new molecular technologies will also require developments in handling, analysis and integration of the resulting data. Such bioinformatics protocols will probably follow swiftly after the emergence of new data, as has been the case in the past (e.g. [174,465,485,486]). A deep understanding of the underlying principles and limitations with which data is generated, informs on the expected outcomes of analyses. Although time-consuming, systematic inspection at all steps of automated data processing and analysis pipelines remains critical for obtaining trustworthy results. For instance, during our comparative evolutionary analyses of primate genomes, intensive manual curation weeded out many artefacts, which were abundant despite a highly stringent analysis pipeline (Chapter 5). Although large-scale sequence analyses are nowadays often considered routine exercises, even comparisons of species that are relatively closely related such as primates still requires vigilance [339,487]. Besides improving analysis outcome, quality control efforts provide insights into the types of challenges encountered not only in studies of positive selection, but in comparative genomics and other fields in general.

**Computational and experimental approaches are complementary equals**

Experimental and computational work are complementary in modern biology, with both approaches gaining value when combined with the other (e.g. [10,11,17], Figure 1 in Chapter 1). Experiments are important for computational approaches for many reasons, including (i) to measure the initial data required for analyses, (ii) to validate predictions, and (iii) to characterize molecular phenomena in detail once new hypotheses have been established. In our experience the success and value of computational analyses depend to a large extent on the inclusion of validation experiments. Much of the work in this thesis has been part of excellent wet-lab collaborations, with our studies either having led to new experiments from an early stage (Chapters 2, 3 and 7) or generating insights through data analysis at a later stage (Chapters 4 and 6). Chapters in which bioinformatics work was followed by experiments are generally more conclusive than chapters for which we currently have no extensive validation, such as our predictions for OXPHOS regulators (Chapter 6) and our positive selection-based analyses of virus-host interactions (Chapter 5). Just like early involvement of computational biologists improves the design and data resulting from large-scale experimental studies, the best way to ensure that computation-driven studies are actually used and tested is by involving specialists such as immunologists, virologists, geneticists, molecular and cell biologists, right from the start.

It is clear that experimental work will remain an integral part of systems biology studies. At the same time, carefully designed and executed computational studies, such as screens for components of molecular systems, are potentially equal or better descriptors of reality than experiments. For instance, integrated biological models can be much more powerful than individual datasets (e.g. [11,52,177,179], Chapters 2 and 7). Yet, in the evaluation of hybrid computational and experimental studies, it sometimes appears to happen that only the experimental outcomes are considered the ‘real’ evidence, with the computational work, probably unconsciously, disregarded as a theoretical model (e.g. [488,489]). One could argue that the two are not much different, however, given that
proper computational studies are firmly based on experimental data in the first place and that experiments are ultimately also a model of sorts. It is likely that the lack of critical evaluation of the computational aspects of hybrid studies at least partly relates to (i) the rapid development of computational sciences, which has created an educational challenge [467], and (ii) the suboptimal state of review systems for assessing interdisciplinary research [488-490]. In order to maximize knowledge generation we need evidence from different fields and study approaches, both computational and experimental. Scientists should therefore continue to aspire a multidisciplinary, holistic approach to biology.

**Biological and clinical relevance of the research**

This closing section will briefly discuss the biological implications and potential clinical applications of our studies. The concepts, techniques and data types applied here were used to study specific model systems, but they are in fact generic and potentially applicable to all biological systems. Most of the work has focused on fundamental research of the innate immune system. Viruses and other pathogens are a major cause of human disease. For example HIV (AIDS), influenza (Spanish, Mexican flu), SARS coronavirus, Ebola and Zika virus have caused epidemic infection outbreaks [491]. Our systems-level characterization of the RLR pathway and evolution-guided investigation of the virus-host interaction have contributed to deciphering pathogen recognition and viral interference of the immune response. Our genetic analyses of patients have revealed genes influencing susceptibility to viral and fungal infections, contributing to better understanding of the interplay between host genetics and environmental presence of pathogens. We have made the data generated in our studies readily available as resources for use by other scientists, including the validated compendia of innate antiviral pathway components and ciliary genes (CiliaCarta), predicted OXPHOS transcription regulators, and positively selected genes and codons in primates. In addition, our work has featured in an article for the general public [492].

Although vaccines and antivirals exist for preventing and treating a select number of viral infections, therapeutics are lacking for many types of viruses. Fundamental understanding of the pathogen recognition mechanisms of our immune system may lead to novel therapeutic strategies, including drugs, vaccines, and antibodies targeting virus-host interactions. For instance, characterization of viral interactions with the immune system has suggested the potential of vaccines derived from viruses lacking functional interferon antagonists [6,70]. Pathogen-associated molecules such as specific RNAs and CpG-containing DNA that induce type I interferons are already used as vaccine adjuvants [7]. Knowledge of immune pathways should guide further developments of adjuvants and other molecules activating the innate immune response. This may also reveal targets for suppressing autoimmunity [203]. The immune system is also increasingly recognized to play a fundamental role in the development and treatment of cancers, with innate immune signaling now emerging as an important but poorly understood determinant [493]. As is the case for most scientific results of fundamental nature, additional studies based on the discoveries by others and ourselves are required to investigate the possibilities outlined here.
References


Samenvatting in het Nederlands

Levende wezens (organismen) bestaan uit microscopisch kleine cellen; het menselijk lichaam bevat er ongeveer 37 biljoen (37 x 1012). Cellen zijn op hun beurt opgebouwd uit grote aantallen van uiteenlopende moleculen. Moleculaire biologie bestudeert voornamelijk eiwitten (proteïnen) en nucleïnezuren, zoals het DNA (ons ‘genoom’). Het genoom is vrijwel identiek in elke cel en bevat ongeveer 20 duizend genen die coderen voor eiwitten. Interacties tussen die genen en eiwitten leiden tot complexe biologische systemen en processen met specifieke functies. Voorbeelden zijn eiwitcomplexen (bv. het ribosoom, dat eiwitten produceert), signaleringsnetwerken (bv. signalen die waarschuwen voor infectie of aanzetten tot celdeling), celstructuren en organellen (bv. het cilium, een sensor aan de buitenkant van de cel, of het mitochondrioon, dat zorgt voor de energiehuishouding van de cel), en groepen van genen waarin mutaties leiden tot vergelijkbare aandoeningen (bv. virus vatbaarheid of stofwisselingsziekten). Fundamenteel onderzoek naar zulke moleculaire systemen is niet alleen belangrijk om te begrijpen hoe gezonde organismen werken, maar levert ook essentiële kennis op over ziekten. Op die manier kunnen ziekten beter worden herkend, voorkomen en behandeld, bijvoorbeeld door middel van nieuwe vaccins en medicijnen.

Modern biologisch onderzoek produceert dankzij snelle technologische vooruitgang steeds meer en betere experimentele gegevens (data) die de moleculaire inhoud van cellen beschrijft. Voorbeelden van zulke ‘genomics’ data zijn genoomsequenties van verschillende organismen, netwerken van interacties tussen eiwitten en genexpressiepatronen. Het omzetten van al die data in biologische inzichten en kennis vereist innovatieve benaderingen met een sterke computationele biologie en bioinformatica component, in multidisciplinaire samenwerkingen met onder meer immunologen, genetici, artsen, virologen en celbiologen.

Het onderzoek in dit proefschrift is gericht op het ontdekken van nieuwe onderdelen van complexe moleculaire systemen, vanuit een computationeel perspectief: Welke genen zijn betrokken bij welke systemen in de cel? De basis van onze studies ligt in het systematisch analyseren en combineren van grootschalige moleculaire data. We presenteren zes studies, gericht op drie biomedische onderwerpen: (i) infectieziekten – specifiek de immuunrespons tegen virussen en schimmels, (ii) stofwisselingsaandoeningen – de mitochondriële energieproductie, en (iii) ciliopathieën – afwijkingen van het cilium.

**Hoofdstuk 1** introduceert het potentieel van grootschalige genomics data om moleculaire, cellulaire en ziektesystemen beter te beschrijven. We geven een uitgebreid overzicht van recente technologische vooruitgangen die hebben geleid tot de huidige rijkdom aan moleculaire data. Daarnaast beschrijven we de belangrijkste soorten data, methodes om deze data te analyseren en integreren, en beschouwen we succesvolle voorbeelden van de ‘integrative genomics’ aanpak.

Het aangeboren immuunsysteem vormt een essentiële verdediging tegen infecties door virussen, bacteriën en schimmels. Ons genoom codeert voor verschillende moleculaire systemen die infecties in de cel kunnen herkennen om vervolgens de immuunrespons op gang te brengen. **Hoofdstuk 2** beschrijft de ontdekking van ongeveer 100 genen betrokken bij de afweer tegen RNA virussen waaronder het grieppivirus (influenza), hepatitis C, Ebola en verschillende verkoudheidsvirussen.
Deze genen blijken deel uit te maken van het RIG-I-like receptor (RLR) signaleringsnetwerk dat type I interferonen produceert, belangrijke eiwitten voor de antivirale respons. Door middel van systematische analyses van publiekelijk beschikbare genomics data ontdekten we tien kenmerken van het RLR systeem, waaronder interacties met virale eiwitten, versnelde evolutie en regulatie door specifieke transcriptiefactoren (eiwitten die aan het DNA binden om genexpressie te reguleren). Integratie van de data met behulp van Bayesiaanse statistiek leverde voor elk gen een waarschijnlijkheidsscore op voor betrokkenheid bij het RLR systeem. Dit resulteerde in voorspellingen voor bijna 200 nieuwe RLR componenten. De helft daarvan werden bevestigd door computationele en experimentele validatie methodes.

Genetische variatie beïnvloedt de individuele vatbaarheid voor infecties en autoimmunité. Hoofdstuk 3 onthult een onverwachte rol voor het virus receptoreiwit MDA5 in de immuunrespons tegen schimmelinfeccties. De eerste tekenen van een functie voor MDA5 in de afweer tegen schimmels vonden we in genexpressie data van macrofagen, een type witte bloedcel. Daarna ontdekten we dat specifieke genetische varianten van MDA5 hoofdzakelijk voorkomen in patiënten met een hoge vatbaarheid voor bloedinfecties met *Candida* schimmelsoorten, welke normaal gesproken onschuldig zijn. Onze bevindingen werden ondersteund door verdere immunologische experimenten.

Hoofdstuk 4 onderzoekt drie patiënten die lijden onder ernstige infecties met het herpes simplexvirus (HSV-2). Sequencing van de eiwit-coderende regio’s van hun genomen (exoom sequencing) identificeerde per patiënt honderden genetische varianten. Om de mogelijke ziekte- veroorzakende varianten te ontdekken onderzochten we genen die betrokken zijn bij de antivirale immuniteit, waaronder de RLR genen uit Hoofdstuk 2. Een variant in het *ZBTB25* gen bleek vervolgens specifiek voor te komen in alle aangedane familieleden van één van de patiënten met HSV-2 infectieproblemen, maar niet in familieleden zonder infectieproblemen. Vervolgonderzoek moet uitwijzen of deze variant inderdaad de infectiegevoeligheid veroorzaakt.

Virussen hebben een uiteenlopend arsenaal aan eiwitten die de immuunrespons van hun gastheer belemmeren. Als gevolg hiervan ondergaan virale genen en specifieke menselijke genen constante verandering. Die menselijke genen veranderen snel om de virus-eiwitten onschadelijk te maken, en virale genen veranderen snel om weer schade aan te kunnen richten. Hoofdstuk 5 bestudeert deze virus-mens interactie vanuit een evolutionair perspectief met behulp van vergelijkende analyse van negen primaatgenomen, die samen ongeveer 50 miljoen jaar van evolutie beschrijven. Hiervoor ontwikkelden we eerst een nieuwe methode om versnelde evolutie als gevolg van positieve selectie te detecteren. We vonden bewijs voor positieve selectie in 3% van de genen in de mens, voornamelijk met immuun-gerelateerde functies. Aan de hand van expressie- en interactiedata konden we vervolgens nieuwe kandidaat genen voorstellen die in genetische conflicten zijn verwikkeld met virussen. Enkele positief geselecteerde posities in die genen lijken daarnaast, op basis van eiwitstructuren, belangrijk te zijn voor de interactie tussen virussen (onder andere mazelen en HIV) en de receptoren waarmee deze virussen cellen infecteren.

Mitochondriële aandoeningen, in het bijzonder defecten van het oxidatieve fosforyleringsysteem (OXPHOS), leiden tot problemen met de energiestofwisseling. Om een beter begrip te krijgen
van de moleculaire basis van deze ziekten onderzoekt **Hoofdstuk 6** hoe het OXPHOS systeem wordt gereguleerd. Daarvoor vergeleken we genexpressie profielen van patiënt cellen met een defect OXPHOS met profielen van gezonde cellen. Analyses van co-expressie patronen en bindingslocaties van transcriptiefactoren op het DNA lieten verschillen zien in regulatie tussen structurele componenten van het OXPHOS systeem en genen die speciaal zorgen voor de opbouw van dat systeem. Onze data suggereren ook enkele nieuwe transcriptiefactoren die een rol spelen in OXPHOS defecten.

**Hoofdstuk 7** past de data integratietechnieken van Hoofdstuk 2 opnieuw toe, ditmaal om de totale verzameling van genen te identificeren dat betrokken is bij het cilium. Het cilium is een essentieel organel dat als een trilhaar uit de menselijke cel steekt. Het is betrokken bij een breed scala aan genetische aandoeningen (ciliopathieën). Als onderdeel van een grote samenwerking hebben wij voor deze studie zeven genomics, proteomics, transcriptomics en evolutionaire datasets geconstrueerd en grondig geanalyseerd. Systematische integratie van deze data leverde genoom-brede scores op en resulteerde in 285 hoogstwaarschijnlijk nieuwe cilium genen. Onafhankelijke experimenten door onze samenwerkingspartners bevestigden een groot deel van de voorspellingen, waaronder een rol in het cilium voor het OSCP1 gen. Ons werk vormt de basis voor CiliaCarta, een uitgebreid compendium van 836 cilium genen.

**Hoofdstuk 8** sluit dit proefschrift af met een algemene discussie en plaatst het gepresenteerde onderzoek in een breder perspectief. We bespreken onder andere de volgende standpunten. (i) De moderne rijkdom aan genomics data gecombineerd met de juiste studiemethoden verschaf grote mogelijkheden om de werking van moleculaire systemen te achterhalen. De gebruikte data en technieken hebben een brede toepasbaarheid die verder gaat dan de hier bestudeerde systemen. (ii) Toekomstig biologisch onderzoek zal, zoals in het verleden, met name worden gedreven door nieuwe technologische ontwikkelingen die nog meer en betere (grootschalige) moleculaire data opleveren. Dit zal nieuwe kansen bieden om moleculaire systemen systematisch te beschrijven. (iii) Recycling en recombinatie van bestaande data vanuit nieuwe invalshoeken is kosteneffectief en van groot belang om biologische kennis uit publieke data te maximaliseren. (iv) Experimentele en computационelle benaderingen vullen elkaar aan en beide profiteren van intensieve samenwerking. (v) Onze studies hebben nieuwe inzichten verschaft in, onder andere, de cellulaire herkenning van virussen, de virus-gastheer interactie en genetische factoren die de vatbaarheid voor schimmel- en virusinfecties beïnvloeden. Verder onderzoek zal uitwijzen of onze resultaten nieuwe aanknopingspunten bieden voor de ontwikkeling van bijvoorbeeld antivirale medicijnen, stoffen die de immuunrespons versterken (adjuvanten) en vaccins.
Summary

Organisms consist of microscopic cells; the human body contains about 37 trillion of them (37 \times 10^{12}). In turn, individual cells are made up of large numbers of diverse molecules. Molecular biology primarily studies proteins and nucleic acids, such as the DNA (our ‘genome’). The genome is virtually identical in every cell and contains about 20 thousand genes coding for proteins. Interactions between these genes and proteins give rise to complex biological systems and processes with specific functions. Such systems include protein complexes (e.g. the ribosome, which produces proteins), signaling networks (e.g. signals that indicate infection or initiate cell division), cellular structures and organelles (e.g. the cilium, a sensor on the outside of the cell, or the mitochondrion, which generates energy for the cell), and groups of genes in which mutations cause similar diseases (e.g. susceptibility to infections or metabolic diseases). Fundamental research into such molecular systems is not only important to understand how healthy organisms work, but also provides essential knowledge about disease. This means diseases can be diagnosed, prevented and treated in a better way, for example with new vaccines and drugs.

As a result of rapid technological developments, modern biological research produces increasingly more and better experimental information (data) describing the molecules of cells. These ‘genomics’ data include for instance genome sequences of various organisms, networks of interactions between proteins, and gene expression patterns. Transforming all those data into biological insights and knowledge requires innovative approaches with a strong computational biology and bioinformatics component, as well as multidisciplinary collaborations with immunologists, geneticists, medical doctors, virologists, and cell biologists.

The research in this thesis is focused on discovering new components of complex molecular systems, from a computational perspective: Which genes are involved in which systems in the cell? Our studies are based on the systematic analysis and combination of large-scale molecular data. We present six studies that investigate three biomedical themes: (i) infectious diseases – specifically the immune response against viruses and fungi, (ii) metabolic disorders – the mitochondrial energy production, and (iii) ciliopathies – disorders of the cilium.

Chapter 1 introduces the potential of large-scale genomics data for the comprehensive description of molecular, cellular and disease systems. We provide a broad overview of recent technological advances that have led to the current wealth of molecular data. In addition, we describe the most important data types, methods to analyze and integrate these data, and we consider successful examples of the ‘integrative genomics’ approach.

The innate immune system forms the first line of defense against infections by viruses, bacteria and fungi. Our genomes encode several molecular systems that are able to recognize infections in the cell to trigger an immune response. Chapter 2 describes the discovery of about 100 genes involved in the defense against RNA viruses, such as hepatitis C, Ebola, the flu (influenza), and different common cold viruses. These genes turn out to be part of the RIG-I-like receptor (RLR) signaling network that produces type I interferons, which are important proteins for initiating the antiviral response. We identified ten characteristics of the RLR system through systematic analysis
of publicly available genomics data, including interactions with viral proteins, rapid evolution, and regulation by specific transcription factors (proteins that bind the DNA to control gene expression). Integration of the data using Bayesian statistics provided a probability score for each gene, representing the likelihood for involvement in the RLR system. The scores led to predictions for almost 200 novel RLR components. Half of these were confirmed by computational and experimental validation methods.

Genetic variation influences the individual susceptibility to infections and autoimmunity. Chapter 3 reveals an unexpected role for the virus receptor protein MDA5 in the immune response against fungal infections. We found the first signs of a function for MDA5 in antifungal immunity in gene expression data of macrophages, a type of white blood cell. We then identified specific genetic variants in MDA5 that predominantly occur in patients susceptible to blood infections with *Candida* fungi, which are normally harmless. Our findings were supported by further immunological experiments.

Chapter 4 investigates three patients suffering from severe infections caused by herpes simplex virus (HSV-2). Sequencing of the protein-coding regions of their genomes (exome sequencing) identified hundreds of genetic variants in each patient. To discover the potential disease-causing variants, we explored genes that are involved in antiviral immunity, including the RLR genes from Chapter 2. A variant in the *ZBTB25* gene was subsequently shown to occur specifically in all affected family members of one of the patients with HSV-2 infection problems, but not in family members without infection problems. Additional studies are required to determine whether this variant indeed causes the susceptibility to infection.

Viruses have a diverse array of proteins that interfere with the host immune response. As a result, viral genes and certain human genes undergo constant change. Those human genes change quickly to neutralize the viral proteins, while the viral genes change quickly to retain their capability to interfere. Chapter 5 studies this virus-human interaction from an evolutionary perspective using comparative analysis of nine primate genomes, which together cover about 50 million years of evolution. For that, we first developed a new method for detecting accelerated evolution caused by positive selection. We found evidence for positive selection in 3% of human genes, mainly those with immune-related functions. Using expression and interaction data, we then proposed novel candidate genes that are in genetic conflicts with viruses. In addition, based on protein structures some of the positively selected positions in those genes appear to be important for the interaction between viruses (including measles and HIV) and the receptors that these viruses use to infect cells.

Mitochondrial disorders, in particular defects of the oxidative phosphorylation (OXPHOS) system, result in energy deficiencies. In order to gain a better understanding of the molecular basis of these diseases, Chapter 6 investigates how the OXPHOS system is regulated. For that, we compared gene expression profiles of OXPHOS-deficient patient cells and healthy cells. Analyses of coexpression patterns and DNA binding sites for transcription factors revealed differences between the regulation of structural components of the OXPHOS system and genes that are specifically
involved in the assembly of that system. Our data also suggest several novel transcription factors that could play a role in OXPHOS defects.

Chapter 7 reapplies the data integration techniques of Chapter 2, this time to identify the total set of genes involved in the cilium. The cilium is an essential organelle protruding from the human cell as a hair-like structure. It is involved in a wide range of genetic conditions (ciliopathies). As part of a large-scale collaboration, in this study we constructed and rigorously analyzed seven genomics, proteomics, transcriptomics and evolutionary datasets. Systematic integration of these data provided genome-wide scores and resulted in 285 novel likely ciliary genes. Independent experiments by our collaborators confirmed many of the predictions and characterized a ciliary role for the \textit{OSCP1} gene. Our work forms the basis of CiliaCarta, a comprehensive compendium of 836 ciliary genes.

Chapter 8 concludes this thesis with a general discussion and places the presented research into a broader perspective. We discuss the following views among other things. (i) The modern wealth of genomics data, combined with the right study methods, provide great opportunities to determine the workings of molecular systems. The data and techniques used in our studies have a broad applicability that goes beyond the systems studied here. (ii) Future biological research will, like in the past, mainly be driven by new technological developments that provide even more and better (large-scale) molecular data. This will offer new opportunities for the systematic description of molecular systems. (iii) Recycling and recombination of existing data from new angles is cost-effective and important for maximizing biological knowledge generated from public data. (iv) Experimental and computational approaches are complementary and both benefit from intensive collaboration. (v) Our studies have provided new insights into, among other things, the detection of viruses in the cell, the virus-host interaction, and the genetic factors that influence the susceptibility to fungal and viral infections. Further research will determine whether our results provide new leads for the development of for instance antiviral drugs, substances that enhance the immune response (adjuvants), and vaccines.
Dankwoord / Acknowledgements

Allerbeste allemaal, Dearest all,

Vele mensen ben ik dankbaar voor hun hulp en ondersteuning – direct of indirect – bij het tot stand komen van mijn proefschrift. Met dit dankwoord wil ik jullie allemaal heel erg bedanken!

Mijn promotor, prof. dr. Martijn Huynen. Martijn, ik kan je niet genoeg bedanken voor je motiverende vertrouwen in mijn kunnen, zowel voorafgaand maar vooral tijdens mijn promotietraject. Bedankt voor de zeldzame vrijheid die je me hebt geboden om mijn eigen onderzoek uit te stippelen, zijprojecten op te zetten, het werk uit mijn master-periode te publiceren, en niet in het minst om mijn proefschrift in alle rust af te ronden. Ook heb ik het zeer gewaardeerd dat je geen onzinnige druk oplegt en dat je deur altijd open staat voor overleg. Kortom, je bent een fijn mens om mee te werken. Je eigenzinnige visie op de wetenschap hoop ik met me mee te dragen!


Madan, thank you for all the opportunities you have offered me not only during my exciting time in Cambridge, but also afterwards. Even though you are incredibly busy, you are always happy to make time when I need your help or connections. You have been instrumental for my development as a scientist and I feel very lucky to have absorbed some of your unparalleled knowledge of computational biology, productivity and analytical thinking.

My awesome CMBI colleagues, I would like to thank all of you for making my time as a PhD student very pleasant and fun. Thanks for the company during our daily lunch walks, which were a much-needed source of fresh air and distraction from the computer screen and work. They almost always took place outside, regardless of dreadful rain, hail, snow, freezing or boiling temperatures! I have really appreciated our countless get-togethers, awesome nights out, gezellige dinners, cool parties at Drift, de vierdaagse, de Kaaij, exciting escape rooms, lekkere beer tastings, and the casual Aesculaaf visits. These events were always relaxing and energizing at the same time, and have been very important for me, especially during the last two years!

John, onze vijf jaar als kamergenoten zitten er sinds kort helaas op. In die tijd heb ik veel geleerd van de zorgvuldige manier waarop je wetenschap bedrijft. We hebben geweldig samengewerkt aan een aantal (aankomende) publicaties; resultaten waar we trots op mogen zijn. Bedankt voor de vele goede gesprekken over de uitdagingen van het promotietraject, de gezellige theepauzes, het gidsen in Dublin, en het wateren van onze essentiële zuurstofbronnen. Heel veel succes terug in Utrecht en blijf die fylogenetische bomen knuffelen!

Rob, na onze gezamenlijke studietijd was je komst naar onze onderzoeksgroep een welkome verrassing. Onze lunchwandelingen hebben tal van nuttige (en nutteloze) inzichten en ideeën
opgeleverd over onze levens en projecten. Het waren de voorlopers van wat ondertussen een beroemde/beruchte CMBI-traditie genoemd mag worden. Naast een messcherp denkvermogen bezit je het vermogen om de hele dag door brood en groente te eten, behalve tijdens de gangbare pauzes, en irriteren/amuseer je velen met je voorliefde voor gekke liedjes (Mickey Mouse, Shiva, ...). Het blijft prachtig om mee te maken hoe je nietsvermoedende doelwitten aan uitgebreide ondervragingen onderwerpt. Je recente publicatiesucces is dik verdiend!

Wouter, onze lange paden van studie tot promotie komen ‘aardig’ overeen en hebben elkaar vaak gekruist. Ik herinner me nog goed ons gezamenlijke gesprek met Gert op het CMBI tijdens onze bachelor stage: het bleek letterlijk en figuurlijk onze landing vanaf het lab naar de bioinformatica. Dat we uiteindelijk binnen een week van elkaar zullen promoveren is gaaf en toepasselijk. Bedankt voor alle leuke eet/bier/film/concertavonden en gesprekken over sport/muziek/wetenschap/onzin. Je bent een echte klasbak en ik heb altijd gretig gebruikgemaakt van je kennis en kunde. Het is een grote eer om je paranimf te mogen zijn!

Lisette, you really brought refreshing new life into the office with your arrival last year. You are great company and in a short time we became good friends, as well as running and fruit break buddies. I am glad I have been able to offer you advice on your projects. The speed with which you are learning and applying computational biology (en Nederlands) is quite incredible. Het gaat zeker goedkomen met jou! Lass mal die Tassen im Schrank; Tschüssi und viel Glück.

Dei, I really enjoyed our many long afternoon ‘breaks’ to brainstorm about our projects, new results, and the best ways to visualize data. Even more fun was talking to you about sports, exchanging R wizardry, your crazy knowledge about plants (woah, melons are definitely cucumbers in disguise), and the Spanish lessons that, as I found out on La Palma, almost exclusively covered useless words. Muchas gracias, mi compañero de trabajo!

Carolien, het is ongelofelijk maar waar: voor jouw komst naar het CMBI was ik nog nooit in de Aesculaaf geweest! Ondertussen gaan we regelmatig met z’n allen op vrijdagmiddag (en op andere dagen..) spontaan wat drinken; altijd gezellig en ontspannend. Epidemiologen blijken dus toch nuttig. Ik vind het knap hoe je jezelf hebt afbeuld om terug te komen van blessures en weer te kunnen voetballen.

Jon, het is maar goed dat je vanuit Engeland naar Nederland bent gekomen, want hier is het natuurlijk veel minder koud en regenachtig... Jouw aanwezigheid heeft ervoor gezorgd dat ik niet alleen maar met wetenschappers hoefde te praten. Ik was sceptisch over je blog/bitcoin/wale/cryptocurrency-avonturen, maar je hebt me zeker overtuigd toen je een rondje gaf van je verdiende euro’s!

Wout, met jou in de buurt escaleert alles gegarandeerd gierend uit de klauwen. Je bent een onuitputtelijke bron van slechte grappen, geniale woordspelingen en slap geouwehoer. Ik heb zelden zó veel, zó hard gelachen als in de periode dat we kamergenoten waren. Ok, en dooorrrrrr!

Taco, je bent dan wel experimenteel bioloog en PI, maar denk maar niet dat je daarom geen nerd en (jazeker: en) geek bent. Bedankt voor je gezelschap en de nuttige gesprekken over cover letters, sollicitatiebrieven en wetenschapspolitiek tijdens onze pre-voetbal eetafspraken. Stay strong man!

Xiaowen, your unusual worldviews are always interesting to talk about! It is crazy that Wout, Wouter, you and me will all defend our theses within one month of each other. Joanna, thanks
for the board game nights and the sarcasm. Kuntal, good luck with your own thesis! Laurens, je weergaloze kennis van speciaalbieren heeft je snel popular gemaakt. Samen met de andere kenners Wouter, Jon en John (s eigen brouwsels) hebben we mooie proeverijen gehad. Selma, succes met je laboronderzoek; vergeet niet gebruik te maken van bioinformatica! Sergio, your realism and thorough approach to science are very useful to remember from time to time. Bas, Radek, Daniel, Richard, I have learned a lot from your suggestions and comments during group meetings and literature discussions; definitely not just during my own presentations. Isabel, Martin, Philip, thanks for your advice during the initial stages of my PhD projects. Sacha, Tom, bedankt voor jullie bamics-gezelschap. Arthur, Hanka, en Celia, bedankt voor de IT hulp, yasara tricks, en buitenlandconnecties. Barbara, bedankt voor het organiseren van de afdelingsuitjes, etentjes en borrels, en je hulp met allerlei randzaken die ervoor zorgt dat we zoveel mogelijk tijd hebben voor wetenschap.

Ik heb de afgelopen jaren veel plezier en voldoening gehaald uit verschillende samenwerkingen. In Utrecht, Frank, bedankt voor het enthousiasme dat je vanaf het begin in ons project hebt getoond. Je scherpe wet-lab/virologische blik op onze resultaten was telkens erg verhelderend en ik heb veel opgestoken van de constructive commentaren. Qian en Martijn, bedankt voor jullie geweldige experimentele werk. Ik vind het nog steeds ontzettend gaaf dat jullie mij mijn voorspellingen écht getest hebben! In Nijmegen bij interne geneeskunde, Mihai, bedankt voor je waardevolle medewerking aan drie verschillende hoofdstukken in dit proefschrift en de bereidheid om jullie data aan ons ter beschikking te stellen. Martin, het was altijd prima om de wandeling naar jullie maffe bunker te maken. Samen hebben we het maximale uit jullie data geperst en een verrassende ontdekking gedaan. Tot in de Aesculaaf! Bij kindergeneeskunde, Rutger, het zag er naar uit dat jouw expressiedata ongebruikt zou blijven voordat ik het op mijn bordje kreeg. Ik vind dat we er samen in korte tijd toch een mooi paper van gemaakt hebben. Bij humane genetica, Peer en Alex, erg leuk dat jullie mij benaderden om mee te kijken naar jullie patiëntendata, zodat ik mijn kennis voor een klinische doel in kon zetten! In Rotterdam, Arno, bedankt voor de enthousiaste discussies over onze projecten en de vele keren dat ik ons werk heb mogen presenteren op de Virgo bijeenkomsten. Jammer dat je niet kan opponeren bij mijn verdediging!

Beste vrienden. Maarten, het was erg tof om je overal tegen te komen, op school in Waalwijk, studeren in Nijmegen en Engeland, en promoveren in Nijmegen. Je oprechte verwondering over random gebeurtenissen en fenomenen is super aanstekelijk en vermakelijk! Bedankt voor de vele diepe en luchtige gesprekken, bijvoorbeeld tijdens etentjes met Rob, en voor het LAN-plezier. Veel plezier in het onderwijs, of waar je ook terecht komt! Marco, onze vriendschap duurt al een kwart eeuw en dat is cool. Studievrienden van JC de B*, onze eigen jaarclub waarvoor we geen studievereniging nodig hadden; Frank, Bart, Marco, Jos, Bram, Robert-Jan, Rob, Maarten, bedankt voor de avondjes lol en slap geouwehoer onder het genot van vele biertjes. De reisjes naar Bratislava, Praag, Londen en Budapest zijn legendarisch en we zullen nog vaak hard lachen bij het ophalen van die herinneringen. Het ga jullie allemaal goed! Marleen, de late avondjes en Kroatië waren super gezellig! Frederik, bedankt voor de diepzinnige gesprekken. Buurtjes, Marco en Irene, bedankt voor de goede zorgen toen ik in de prak lag met maag en enkel! Christian, thanks for adopting our football team when our own group had a chronic lack of players. Thursday
evenings have been a great source of exercise and fun together with Taco, Wouter, Dei, Joanna, Ran, Thomas, Maarten and many others.


Mijn lieve zus, Eline, bedankt voor onze fijne band en je ontelbare verhalen, grappen en grollen. Je bent een voorbeeld voor me in hoe je klaarstaat voor anderen; vergeet niet ook aan jezelf te denken! Doe Hendrik Ido de groeten.

Mijn geweldige ouders, lieve Mam en Pap, bedankt voor jullie onvoorwaardelijke steun, goede raad en aanmoedigingen bij alles wat ik ooit wilde doen. Bedankt ook voor de hulp bij al mijn verhuizingen, de mooie bergwandeling- en wintersportvakanties, en de oneindige hoeveelheid eten die ik heb opgemaakt! Bij jullie kom ik altijd tot rust en voelt het echt als thuiskomen. Ik kan me geen betere ouders wensen!

Lieve Eline, jouw positiviteit, warmte en zorgzaamheid zijn onuitputtelijk. Je weet precies wanneer ik afleiding nodig heb en wanneer je me moet laten. Ik ben zeker niet altijd de makkelijkste; bedankt voor je inspanningen om me te helpen ontspannen 😊. Onze vakanties naar de natuur zijn erg belangrijk voor me. Samen met jou erop uit trekken is het mooiste wat er is. Ik kijk uit naar ons volgende avontuur!

– Robin

Nijmegen, september 2016


* Co-first authors
Curriculum Vitae

Robin van der Lee was born on August 27th 1988 in Waalwijk, The Netherlands. He went to the Dr. Mollercollege in Waalwijk and received his high school VWO (Gymnasium) diploma cum laude in 2006, having completed the curricula Nature, Health & Technology. He then moved to Nijmegen to study Molecular Life Sciences at the Radboud University, where he received a broad training in chemistry, molecular biology and mechanisms of disease. Having completed various research internships, he received both his Bachelor’s degree (in 2009) and Master’s degree (in 2011) cum laude. During his Master’s studies, he specialized in computational biology and bioinformatics. He worked in protein structure bioinformatics on the prediction of protein stability changes upon mutation, with Prof. dr. Gert Vriend, head of the Centre for Molecular and Biomolecular Informatics (CMBI) at the Radboud university medical center in Nijmegen. He was then awarded the European Erasmus scholarship for an 8-month visit to the systems biology and regulatory genomics group of Dr. M. Madan Babu at the MRC Laboratory of Molecular Biology in Cambridge (UK). There he studied various aspects of intrinsically disordered proteins (IDPs), which led to publications on the classification of IDPs and the role of IDPs in protein degradation. After successfully completing his Master’s degree, he returned to the CMBI in 2011 to start his PhD studies in the comparative genomics group of Prof. dr. Martijn Huynen. As part of the Dutch virus research collaboration VIRGO, Robin investigated a wide range of genomics data to understand various molecular systems including innate antiviral immunity, mitochondria and the cilium. The results of his research are described in this thesis.