Host and pathogen genomics of severe pediatric infections

THÈSE N^{O} 6656 (2016)

PRÉSENTÉE LE 27 AVRIL 2016

À LA FACULTÉ DES SCIENCES DE LA VIE
GROUPE FELLAY
PROGRAMME DOCTORAL EN BIOTECHNOLOGIE ET GÉNIE BIOLOGIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Samira ASGARI

acceptée sur proposition du jury:

Prof. B. Deplancke, président du jury
Prof. J. Fellay, directeur de thèse
Prof. J.-L. Casanova, rapporteur
Prof. P. McLaren, rapporteur
Dr J. Rougemont, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2016

Acknowledgements

A great many people have contributed to the completion of this dissertation. I owe my gratitude to every one of these people who have made my graduate experience a period that I will cherish forever.

Foremost, I would like to thank my advisor, Dr. Jacques Fellay, whose expertise, knowledge and thoughtful feedback have guided me through my graduate studies. I appreciate the freedom he has given me to explore on my own, and the guidance he provided when my progress faltered. Jacques' encouragement, understanding and patience added considerably to my overall experience.

I would like to extend my thanks to Dr. Didier Trono, my mentor. Didier's insightful advice was pivotal in my decision to follow a career in computational biology and his meticulous suggestions have always helped me to clarify my ideas.

I would like to acknowledge Dr. Luregn Schlapbach for his crucial role in the recruitment of patients, Dr. Dominique Garcin for his indispensable role in experimental validation of our findings and Dr. Caroline Tapparel for her expert input in designing the experiments.

My sincere thanks also go to our patients and their families. Without their trust and participation this work would never have been possible.

I would like to thank many past and current colleagues who, directly or indirectly, contributed to completion of this work. In particular, I would like to thank my colleagues in Fellay lab. Their thought-provoking discussions helped me to focus my ideas, their constructive criticisms held me to the highest of standards at all stages of my research, and their humour and conviviality made the day-to-day work experience filled with joy. This is by necessity an incomplete list, in no particular order: Marisa Marciano wynn, Paul McLaren, Thomas Junier, Istvan Bartha, Christian Hammer, Nimisha Chaturvedi, Alessandro Borghesi, Petar Scepanovic, Christian Thorball, Ana Bittencourt Piccini, Caroline Perraudin, Jennifer Makovkina, Margalida Rotger and Antonio Rausell.

I'm blessed with numerous extraordinary friends to whom I owe some of the happiest memories of my life. Their support helped me overcome the moments of crisis and their craziness helped me stay sane in difficulties. My deepest appreciation goes, in no particular order, to: Lionel, Amin, Somayyeh, Mahdi, Farhang, Victoria, Arefeh, Azadeh, Hanieh, Mahdieh and Vahideh.

Finally, I owe my deepest gratitude to my family. They have been a constant source of love, support and strength throughout my life. My heart-felt gratitude goes to my parents, my sister and my brother. This dissertation is dedicated to them.

Abstract

Infectious diseases are among the leading causes of human morbidity and mortality, with the greatest burden felt in the pediatric population. For any infectious disease, only a fraction of the exposed individuals develop clinical symptoms. These inter-individual differences can be due to variation in pathogen virulence or in host susceptibility. The recent advent of high-throughput sequencing (HTS) technology has enabled studies of both human and pathogen genetic factors that have the potential to influence infectious diseases pathogenesis and alter clinical presentation.

In this thesis, I present a set of genomic studies that used HTS to dissect the genetic basis of life-threatening infections with *Pseudomonas aeruginosa* (*P. aeruginosa*) and respiratory syncytial virus (RSV). This work provides conclusive evidence for the role of rare human genetic variants in susceptibility to life-threatening *P. aeruginosa* and RSV infections in previously healthy children. Furthermore, in an attempt to determine the role of viral genetic factors in severe presentations of RSV infection, I established a framework for exploring RSV genetic variation using HTS technology and bioinformatic analysis.

Together, theses studies demonstrate that current genomic technology, bioinformatic analysis and functional follow-up have the potential to give us novel insight into the molecular basis of host-pathogen interactions and infectious disease pathogenesis.

Keywords: Infectious diseases, *Pseudomonas aeruginosa*, respiratory syncytial virus, pediatric population, human genomics, viral genomics, high-throughput sequencing, primary immunodeficiency

Résumé

Les maladies infectieuses comptent parmi les causes principales de morbidité et de mortalité chez l’homme, et la population pédiatrique est particulièrement touchée. Pour une maladie infectieuse donnée, seule une fraction des individus exposés développent des symptômes cliniques. Ces différences interindividuelles peuvent être dues à des variations dans la virulence du pathogène ou dans la susceptibilité de l'hôte. Les récents développements dans le domaine du séquençage à haut débit ont ouvert la voie à l'étude des facteurs génétiques de l'humain comme des pathogènes - des facteurs qui ont le potentiel d'influencer la pathogenèse des maladies infectieuses ou d'en modifier la présentation clinique.

Dans cette thèse, je présente une série d'études reposant sur le séquençage à haut débit, qui ont pour objectif d’élucider les bases génétiques d'infections potentiellement mortelles dues à *Pseudomonas aeruginosa* (*P. aeruginosa*) et au virus respiratoire syncytial (RSV pour *respiratory syncytial virus*). Ce travail a permis de mettre en évidence le rôle de variations génétiques humaines rares dans la susceptibilité de certains enfants sans facteurs de risque établis de développer une maladie potentiellement mortelle suite à une infection par *P. aeruginosa* ou RSV. D’autre part, j'ai développé un modèle d’exploration des variations génétiques du RSV reposant sur le séquençage à haut débit et l'analyse bioinformatique.

Ces études démontrent que la combinaison des technologies génomiques les plus récentes, d’analyses bioinformatiques et d’un suivi fonctionnel permet de fournir de nouvelles informations sur les bases moléculaires des interactions entre hôte et pathogène, ainsi que sur la pathogenèse des maladies infectieuses.

Mots-clés: maladies infectieuses, *Pseudomonas aeruginosa*, virus respiratoire syncytial, population pédiatrique, génomique humaine, génomique virale, séquençage à haut débit, immunodéficience primaire

Table of contents

Acknowledgements .... **iii**

Abstract .... **v**

Résumé .... **vii**

Table of contents .... **ix**

List of figures .... **xi**

List of tables .... **xii**

Chapter 1 Introduction **13**

1.1 Genetic susceptibility to infectious diseases **13**
1.2 Approaches to gene discovery **13**
1.2.1 Genome-wide linkage and candidate-gene studies 13
1.2.2 Genome-wide association studies (GWAS) 14
1.2.3 High-throughput sequencing (HTS) 16
1.3 Genetic architecture of infectious diseases **20**
1.4 Role of pathogen **22**
1.5 Justification of research **23**
1.6 References **24**

Chapter 2 Exome sequencing reveals primary immunodeficiencies in children with *Pseudomonas aeruginosa* sepsis **35**

2.1 Introduction **35**
2.2 Material and methods **35**
2.2.1 Patients 35
2.2.2 DNA extraction and exome sequencing 35
2.2.3 Alignment of sequencing reads, variant calling and annotation 36
2.2.4 Identification of potential causal variants 36
2.2.5 Gene-annotation enrichment analysis 36
2.3 Results **36**
2.3.1 Clinical presentation and outcomes 36
.... **37**
2.3.2 Exome sequencing 38
2.3.3 Variant analysis 40
2.4 Conclusion and discussion **44**
2.5 References **46**

Chapter 3 Loss-of-function mutations in IFIH1 predispose to severe viral respiratory infections in children **47**

3.1 Inroduction **47**
3.2 Material and methods **47**
3.2.1 Subject recruitment and specimen collection 47
3.2.2 Screening of respiratory viruses 48
3.2.3 Exome sequencing and alignment 48
3.2.4 Variant calling 49

3.2.5 Variant annotation 49
3.2.6 RNA sequencing and alignment 49
3.2.7 Plasmids 49
3.2.8 Transfection & measure of interferon-β promoter activity 50
3.2.9 Viruses 51
3.2.10 Measure of protein stability by pulse chase 51
3.2.11 Recombinant IFIH1 expression 51
3.2.12 Measure of IFIH1 ATPase activity 51
3.2.13 Cell culture and transduction with lentivirus vectors 52
3.2.14 Quantitative RT-PCR 52
3.2.15 Statistical analysis 52
3.3 Results 52
3.3.1 Clinical presentation and outcomes 52
3.3.2 Exome sequencing and alignment of short reads 54
3.3.3 Variant calling and annotation 54
3.3.4 Loss of function variants 55
3.3.5 IFIH1 role in innate immunity 55
3.3.6 Description of *IFIH1* LoF carriers 56
3.3.7 Description of *IFIH1* LoF mutations 56
3.3.8 RNA sequencing and read mapping of IFIH1 LoFs 57
3.3.9 Functional characterization of IFIH1 LoF mutations 59
3.3.10 Role of IFIH1 in controlling respiratory viruses 61
3.4 Conclusion and discussion 63
3.5 References 64

Chapter 4 Deep sequencing and *de novo* assembly of two respiratory syncytial virus genomes isolated from previously healthy children with severe infection . 67
4.1 Introduction 67
4.2 Material and methods 68
4.2.1 Subject recruitment and specimen collection 68
4.2.2 Viral RNA sequencing 68
4.2.3 De novo assembly and analysis 68
4.2.4 Maximum likelihood phylogeny analysis 69
4.2.5 Protein similarity analysis 69
4.3 Results 69
4.3.1 Study subjects 69
4.3.2 Viral RNA sequencing 70
4.3.3 *de novo* assembly and analysis 70
4.3.4 Maximum likelihood tree generation 78
4.3.5 Protein similarity analysis 81
4.4 Conclusion and discussion 85
4.5 References 86

Chapter 5 Conclusion 89

Curriculum Vitae 91

List of figures

Figure 1.1: Relationship between allele frequency and effect size 15

Figure 1.2: A model for genetic architecture of infectious diseases 22

Figure 2.1: Ecthyma grangrensum 38

Figure 2.2: Serum lytic activity. 44

Figure 3.1: Description of *IFIH1* LoF mutations 58

Figure 3.2: Functional characterization of *IFIH1* LoF mutations 60

Figure 3.3: Role of IFIH1 in controlling HRV and RSV infections 62

Figure 4.1: *de novo* assembly of RSV genome 71

Figure 4.2: Maximum likelihood phylogeny tree 80

Figure 4.3: PRI-024 protein similarity dot plot 82

Figure 4.4: PRI-030 protein similarity dot plot 83

List of tables

Table 1.1: Novel infectious diseases genes discovered using high-throughput sequencinh since 2010 16

Table 2.1: Patients' demographic information 37

Table 2.2: BWA alignment metrics 38

Table 2.3: GATK variant calling metrics 39

Table 2.4: SnpEff annotation metrics 40

Table 2.5: Potentially causal variants 41

Table 2.6: Rare nonsynonymous variants in PID genes 41

Table 3.1: Summary of patients' demographic information 53

Table 3.2: Summary of BWA alignment metrics 54

Table 3.3: Variant calling metrics 54

Table 3.4: Completely knocked-out genes 55

Table 3.5: LoF mutations found in *IFIH1* 56

Table 3.6: RNA-sequencing mapping metrics 57

Table 4.1: BLAST analysis for PRI-024 *de novo* assembled contigs 72

Table 4.2: BLAST analysis for PRI-030 *de novo* assembled contigs 76

Table 4.3: Pairwise alignment results 79

Table 4.4: Pairwise comparison of unique amino acid changes 84

Chapter 1 Introduction

1.1 Genetic susceptibility to infectious diseases

Heritable factors have been considered to play a role in infectious diseases for a long time. Tuberculosis and leprosy, for example, were long considered heritable diseases as they were usually clustered within a household (*1*, *2*) . With Pasteur's discovery of microbes, and the observations supporting the communicable nature of infectious diseases, the focus moved from host to pathogen. The role of human genetics in the outcome and clinical features of infectious diseases was highlighted again by multiple observations that exposure to a pathogen, while necessary for infection, is not sufficient, and that not all exposed individuals develop overt disease. The first such observation was Charles Nicolle's discovery of "asymptomatic infections" between 1911 and 1920. Nicole realized that apparently healthy individuals can be carriers of replicating typhus microbes (*Rickettsia* bacteria) in their blood and can transmit the disease while remaining completely asymptomatic (*3*, *4*) . In 1927, the accidental infection of a population with *Mycobacterium tuberculosis* in Lubeck, Germany, caused severe disease and death in some individuals, while others remained unaffected (*5*) . Such observations suggested a potential role for human genetics in susceptibility and outcome of infectious diseases (*6–8*) . More evidence for the importance of genetic variation in infectious disease heterogeneity comes from early twin and adoptee studies. In 1978, Comstock showed that in a cohort of twins with tuberculosis, the disease concordance is significantly higher in monozygotic than in dizygotic twins (*9*) . In a landmark study published in 1988, Sorensen et al. showed that adopted children have a markedly higher risk of premature death form infection if one of their biological parents also died from an infectious disease and concluded: "... premature death in adults has a strong genetic background, especially death due to infections and vascular causes" (*10*) . Together, these early studies strongly support the hypothesis that the host genetic profile plays an important role in individual susceptibility to infection.

1.2 Approaches to gene discovery

1.2.1 Genome-wide linkage and candidate-gene studies

The twin and adoptee studies described above were not designed to provide information on the molecular basis of the observed variability in susceptibility to pathogens. In 1952, Ogden Bruton described the first documented case of a primary immunodeficiency (PID), X-linked agammaglobulinemia (XLA). PIDs are inborn errors of immunity that make affected subjects highly susceptible to infectious agents that rarely cause severe outcomes in the vast majority of patients. Bruton described a boy with recurrent invasive pneumococcal disease and who lacked serum gammaglobulins (*11*, *12*) . It took more than

50 years to discover the gene underlying XLA. By 1990, candidate gene studies and genome-wide linkage analysis studies had become standard approaches to study host genetic susceptibility to infectious diseases. They identified a number of genetic regions associated with susceptibility to several infections, including malaria, HIV, tuberculosis, invasive pneumococcal disease and leprosy (*13*, *14*) . However, a clear limitation of the linkage approach is the requirement to recruit families that are potentially informative about infectious disease phenotypes. In candidate gene studies, the choice of target genes was either based on evidence from animal studies, genome-wide scans and clinical observations, or simply based on known biological relevance. For example, the human leukocyte antigen (HLA) locus on chromosome 6 has been under intense scrutiny (*15–18*) . Candidate gene association studies are limited by a need for a pre-existing hypothesis as well as low level of reproducibility. Despite their limitations, findings of genome-wide linkage studies and candidate gene studies, along with *in vivo* studies using different strains of inbred mice shed some light on the molecular basis of susceptibility to several infectious diseases.

1.2.2 Genome-wide association studies (GWAS)

In the first decade of the twenty-first century, the completion of human genome project along with the HapMap project and the advancement in microarray-based high-throughput technologies made it possible to genotype hundreds of thousands of common polymorphisms in the genome of a large number of study participants. These technological and conceptual advances made it possible for the first time to explore the entire human genome without a need to make assumptions about the location of the causal variant(s). The first GWAS was published in 2005 (*19*) . In 2007, the first GWAS of an infectious disease was performed on human immunodeficiency virus-1 (HIV-1) virus load. In this study, Fellay *et al*. described a series of single nucleotide polymorphisms (SNPs) in the major histocompatibility complex (MHC) region that associated with control of viral load and disease progression in a European population (*20*) . GWAS have now been performed for multiple infectious diseases, and some identified new associations between human genetic variants (mainly SNPs) and the disease phenotype. A large GWAS conducted by the International HIV Controllers Study in 2010, confirmed the previously published associations with HIV-1 control, and reported associations for 6 additional SNPs: two in European and four in African-American populations (*21*) . Another example of GWAS success is the detection of an association between SNPs in the *IL28B* gene region and two phenotypes of hepatitis C virus infection: response to interferon-based treatment and spontaneous clearance (*22*) . In 2009, Zhang *et al.* reported the first GWAS of leprosy susceptibility in 11000 individuals of Chinese ancestry. The study identified multiple strong associations, many of which clustering in innate immune pathways (*23*) . The NHGRI-EBI Catalogue of genome-wide association studies contains 76 studies describing 682 associations in 46 infectious phenotypes that were published between 2007 and April 2015 (http://www.ebi.ac.uk/gwas/home). There are now published GWAS for HIV-1 (*21*, *24–42*) , epstein-barr virus (EBV) (*43*) , human papilloma virus (HPV) (*44*) , tripanosoma cruzi (*45*) , hepatitis C virus (*22*, *46–60*) , tuberculosis (*61–65*) , malaria (*66–69*) , hepatitis B virus (*70–79*) , influenza virus (*80*, *81*) , bacteraemia (*82–84*) , leprosy (*23*, *85*, *86*) , leishmaniasis (*87*) , meningococcal disease (*88*) , shingles (*89*) , dengue fever (*90*) , and Creutzfeldt–Jakob disease (*91*, *92*) .

GWAS have once again confirmed the important role of host genetics in shaping human response to various infections. However, for statistical reasons, the GWAS approach can only be successful for common infectious phenotypes associating with alleles that are relatively common in the population. Indeed, a GWAS is a statistical test of association that tests the null hypothesis: “none of the SNPs in this dataset is associated with the trait of interest” against the alternate hypothesis: “at least one of the SNPs in this dataset associates with the trait of interest”. In a GWAS with a set number of observations the power to reject the null hypothesis when the alternate is true depends on sample size, effect size and allele frequency. The standard to declare a GWAS hit significant at genome-wide level is a combined P value (including ‘initial discovery’ GWAS and replication cohorts) of < 5e–8. For less common infectious disease phenotypes, the required sample size to achieve such significance thresholds is unrealistically large, making it very difficult or even impossible to recruit enough study participants. Additionally, even with large sample sizes, a GWAS will not be able to find novel or rare variants, as SNP arrays focus on previously known genetic polymorphisms with the majority of the variants having a minor allele frequency above 1 to 5% (Figure 1.1). Despite these limitations, GWAS improved our understanding of genotype-phenotype relationships in infectious diseases and in some cases led to the discovery of new genetic association in previously unsuspected regions of genome.

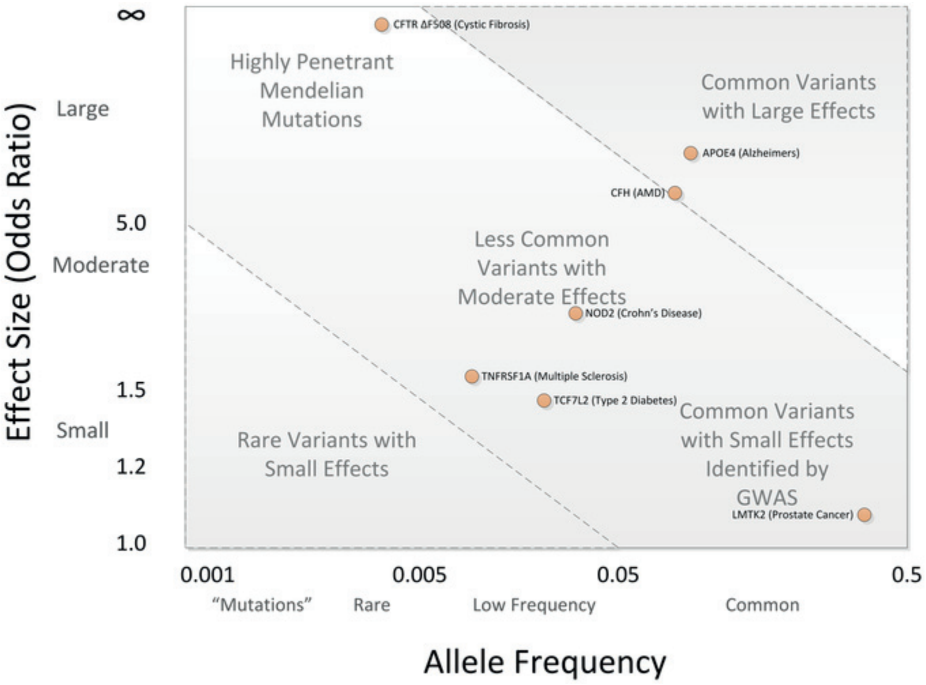

Figure 1.1: Relationship between allele frequency and effect size

Adopted from reference 93; Genome-wide association studies (GWAS) arrays contain variants with allele frequency above 1%. Most GWAS findings are associations between common phenotypes and common alleles with small effect sizes (*93*) .

1.2.3 High-throughput sequencing (HTS)

In 2005, the first high-throughput sequencer became commercially available (the 454 machine, from Life Sciences). The technology used, called pyrosequencing, is based on sequencing-by-synthesis, in which a DNA sequence serves as template and a DNA polymerase is used to extend a sequencing primer by incorporating nucleotides that form a growing sequence complementary to the template. This was the first example of what became known as “next generation sequencing” methods, which used parallelization to produce more reads in a faster and cheaper way (*94*) . The rapid advancement of HTS technologies coupled with the plummeting cost of sequencing allowed the widespread use of this technology in many fields of biology and medicine, including infectious disease biology, during the past five years. HTS in combination with appropriate study design and sampling has made the non *a priori* screening of entire human genomes a reality. A successful example was the use of exome sequencing by Byun *et al.* in 2010 to diagnose a child with fatal classic Kaposi sarcoma (KS), a neoplasm caused by human herpes virus 8 (HHV8). They reported a homozygous splice-site mutation in *STIM1* (chromosome 11 position 4103413, hg19) that was causing immunodeficiency 10 (IMD10) (*95*) . The same year, Bolze *et al.* reported the use of a combination of exome-sequencing and linkage analysis in a large kindred with four affected members to identify a homozygous nonsynonymous mutation in *FADD* as the cause of an autosomal recessive disorder with biological features of autoimmune lymphoproliferative syndrome (ALPS) (*96*) . These two studies were the first example of using HTS in a small number of patients to decipher the genetic basis of rare infectious disorders. Since 2010, there has been an exponential increase in the number of genes known to confer unusual susceptibility to one or more pathogens. Table 1.1 lists these genes discovered since 2010 using HTS.

In brief, recent findings suggest that the use of exome or whole genome sequencing has enormous potential to identify novel variants involved in acquisition or progression of infectious diseases.

Table 1.1: Novel infectious diseases genes discovered using high-throughput sequencinh since 2010

These genes have been identified using exome / genome sequencing approaches or a combination of HTS and linkage study or candidate gene sequencing. AR: Autosomal recessive, AD: Autosomal dominant, XR: X-linked recessive, Ref.: Reference, Inh.: Inheritance, GOF: Gain-of-function, * Evidence from linkage or candidate gene study and evidence from HTS for the role of the gene in the indicated phenotype were published within a few months from each other.

Gene	Mutation	Inh.	Phenotype	Study population	Ref.
FADD	Missense	AR	Severe Infections, autoimmunity	One consanguineous pedigree	(*96*)
MAGT1	Nonsense, frame-shift	XR	Severe viral infections, T cell deficiency, CD4 lymphopenia	Two pedigrees	(*97*)
ZBTB24	Nonsense, missense	AR	B cell deficiency, dysmorphism, Immunodeficiency, Centromeric Instability, and	Nine pedigrees (six consanguineous)	(*98*)

			Facial Anomalies Syndrome Type 2 (ICF2)		
STAT1	Missense (GOF)	AD	Chronic mucocutaneous candidiasis, autoimmune thyroiditis	Multiple pedigrees	(*99*, *100*)
GATA2	Missense, splice site, frame-shift, nonsense	AD	Infections, monocytopenia, B-cell and NK-cell lymphopenia, predisposition to acute myeloid leukemia	Pedigrees and sporadic cases	(*101* – *104*)
RHOH	Nonsense	AR	Various infections, T cell deficiency	One consanguineous pedigree	(*105*)
CARD11	Missense (GOF)	AD	B cell lymphocytosis	One pedigree and one sporadic case	(*106*)
MCM4	Splice site, fame-shift	AR	Adrenal insufficiency, NK cell deficiency	Seven consanguineous pedigree	(*107*, *108*)
POLE	Splice site	AR	Bacterial infection, facial dysmorphism, immunodeficiency, livedo, and short stature (FILS syndrome)	One consanguineous pedigree	(*109*)
TBK1	Missense	AD	Herpes simplex encephalitis	Two sporadic cases	(*110*)
CARD14*	Splice site, missense, in-frame deletion (GOF)	AD	Psoriasis	Six pedigrees	(*111* – *113*)
PLCG2	Missense, large deletion (GOF)	AD	Recurrent infections, low antibody levels, autoimmunity, cold urticaris	Five pedigrees	(*114*, *115*)
PIK3R1	Nonsense	AR	Colitis, absent B cells, agammaglobulinemia	One consanguineous pedigree	(*116*)
ADAR	Missense	AR	Aicardi-Goutières syndrome	Eight pedigrees (two consanguineous)	(*117*)
ISG15	Nonsense	AR	Mycobacterial disorder	Two consanguineous pedigrees	(*118*)
RBCK1	Nonsense, large deletion	AR	Invasive bacterial infections, chronic autoinflammation	Two pedigrees (one consanguineous)	(*119*)
CD27	Nonsense	AR	Impaired T cell-dependent B-cell responses, T-cell dysfunction, EBV infection	One consanguineous pedigree	(*120*)
LRBA*	Missense, nonsense, frame-shift,	AR	Autoimmunity, hypogammaglobulinemia	Six consanguineous pedigree	(*121* –

	large deletion				*123*)
MALT1	Missense	AR	T cell deficiency	One consanguineous pedigree	(*124*)
IL21R	Missense, in-frame deletion	AR	Impaired T and B cells, impaired NK cell cytotoxicity	Two consanguineous pedigrees	(*125*)
CARD11	Deletion, Nonsense	AR	Deficient T cell function, distorted B cell populations	Two consanguineous pedigrees	(*126*, *127*)
TNFRSF4	Missense	AR	Kaposi Sarcoma	One consanguineous pedigree	(*128*)
TTC7A	Splice site, missense	AR	Congenital multiple intestinal atresia	Twelve pedigrees (two consanguineous), three sporadic cases	(*129*, *130*)
TRAF3IP2	Missense	AR	Chronic mucocutaneous candidiasis	One consanguineous pedigree	(*131*)
TCF3	Missense	AD	B cell deficiency, agammaglobulinemia	Eight pedigrees	(*132*)
TNFSF12	Missense	AD	Recurrent infections, B cell deficiency, agammaglobulinemia	One pedigree	(*133*)
NFKB2	Nonsense, frame-shift	AD	Recurrent infections, hypogammaglobulinemia, autoimmune features, adrenal insufficiency	Two pedigrees	(*134*)
PRKCD	Missense, Splice site	AR	Autoimmunity, systemic lupus erythematosus	Three pedigrees (two consanguineous)	(*135*–*137*)
PIK3CD	Missense (GOF)	AD	Recurrent respiratory and ear infections, decreased circulating B and T cells	Eight pedigrees	(*138*)
RPSA	Missense, frame-shift, nonsense	AD	Isolated congenital asplenia (ICA)	Eight pedigrees	(*139*)
VPS45	Missense	AR	Congenital neutropenia, enlarged kidneys	Seven consanguineous pedigrees	(*140*, *141*)
RTEL1	Nonsense, missense, splice region	AR	Hoyeraal–Hreidarsson syndrome, short telomere, hypogammaglobulinemia, cytopenia	Three pedigrees (one consanguineous)	(*142*, *143*)
IKBKB	Nonsense,	AR	Multiple infections,	One	

	frame-shift		hypogammaglobulinemia	consanguineous pedigree, four sporadic cases	(*144*, *145*)
BCL10	Splice site	AR	Combined immunodeficiency with B cell, T cell, and fibroblast defects	One consanguineous pedigree	(*146*)
CTPS1	Splice site	AR	Impaired proliferation of T and B cells t in response to antigen receptor-mediated activation	Five pedigrees	(*147*)
TRNT1	Missense, frame-shift, splice site	AR	B-cell immunodeficiency, periodic fevers, and developmental delay (SIFD)	Eleven pedigrees (three consanguineous) and three sporadic cases	(*148*)
IL21	Missense	AR	Inflammatory bowel disease, reduced numbers of circulating CD19 positive B cells	One consanguineous pedigree	(*149*)
CTLA4	Nonsense, splice site	AD	Recurrent infection, hypogammaglobulinemia, disrupted T and B cell homeostasis	Four pedigrees	(*150*)
TMEM173	Missense (GOF)	AD	Upregulated type I IFN production, autoimmunity	One pedigree	(*151*)
NLRC4	Missense (GOF)	AD	Neonatal-onset enterocolitis, periodic fever, and fatal or near-fatal episodes of autoinflammation	One pedigree	(*152*)
MAP3K14	Missense	AR	Recurrent infections, B cell lymphopenia, hypogammaglobulinemia, decreased NK cells	One consanguineous pedigree	(*153*)
TPP1	Missense, in-frame deletion	AR	Bone marrow failure, Hoyeraal-Hreidarsson syndrome	One pedigree	(*154*)
JAG1	Nonsense. missense	AR	Invasive bacterial infections, severe congenital neutropenia	Two consanguineous pedigrees	(*155*)
INO80	Missense	AR	Defective Immunoglobulin class-switch recombination	Two sporadic cases	(*156*)
STAT4	Missense	AD	Kaposi Sarcoma	One consanguineous pedigree	(*157*)
TADA2A	Missense	AR	polyarteritis nodosa , stroke, hepatosplenomegaly, hypogammaglobulinemia, recurrent infections	Multiple pedigrees	(*158*, *159*)
IFIH1	Missense (GOF)	AD	Upregulated type I IFN production, Aicardi-Goutières syndrome	Seven sporadic cases	(*160*)

IRF7	Missense	AR	Perturbed type I and type III IFN production	One pedigree	(*161*)
DOCK2	Missense, frame-shift	AR	Invasive infections; T cell lymphopenia; impaired T cell, B cell, and natural killer cell function; and defective interferon immunity	Five pedigrees (three consanguineous)	(*162*)
RNF31	Missense	AR	Multiorgan autoinflammation, combined immunodeficiency	One consanguineous pedigree	(*163*)
PGM3	Missense, frame-shift	AR	Recurrent infections, eczema, and increased serum IgE levels	Four consanguineous pedigree	(*164*)
MAP3K9	Nonsense	AR	Susceptibility to severe bacterial infection	One consanguineous pedigree	(*165*)
RORC	Missense, nonsense	AR	Mococutaneous candidiasis and mycobacteriosis	Three consanguineous pedigrees	(*166*)
IRF3	Missense	AD	Herpes simplex encephalitis	16 sporadic cases	(*167*)

1.3 Genetic architecture of infectious diseases

One hundred years after Charles Nicole's description of "asymptomatic infections", the genetic architecture of human susceptibility of most infectious diseases is still unknown. Genetic predisposition to infections can be found on a continuous spectrum from purely Mendelian (caused by fully penetrant alleles of high effect size in a single gene) to polygenic (caused by many variations of low effect size with incomplete penetrance in multiple genes) (*3*) . Till recently, rare infectious phenotypes were thought to cause susceptibility to multiple infections (e.g. most PIDs) and common infectious phenotypes cause susceptibility to one pathogen (e.g. susceptibilities discovered by GWAS). This view is consistent with "rare variant-rare disease and common variant-common disease" hypothesis. This model gained experimental support from identification of rare monogenic variants as the genetic basis of PIDs and from a large number of GWAS showing an association between multiple common variants and common infections. However, this model cannot explain the rare PIDs that predispose to only one infection (or even to only primary infection) with one pathogen (*8*, *168*) .

The first two PIDs conferring predisposition to a single infectious agent to be described were epidermodysplasia verruciformis (EV), and X-linked lymphoproliferative disease (XLP). EV is an extremely rare disorder that predisposes to wart-like skin lesions and skin cancer. It is an interesting, but certainly not unique example of such infectious phenotypes. EV is caused by a specific group of HPV that are harmless in the general population. The genetic origin of EV was suspected since 1933 and its viral etiology was established in 1946 (*169*) . In 2002, recessive mutations in *TMC6* and *TMC8* were described as EV-causing, and two years later EV was added to the international list of PIDs (*170*) . Even though EV as a clinical entity was known even before XLA, described in 1952 as the first

PID, the lack of overt immunological phenotype and susceptibility to only one infection prevented EV to be recognized as a PID at that time (*168*) . Mutations in *SAP* were described in 1998 as the genetic basis of XLP, a rare disease occurring in individual carrying the mutation upon infection with EBV (*171*) . XLP is an example of a Mendelian infectious disease with variable phenotype, as patients can develop hemophagocytosis, lymphoma, or hypogammaglobulinemia (*168*) . Mendelian Susceptibility to Mycobacterial Disease was the first example of Mendelian susceptibility to a coherent group of pathogens, and the genetic etiology of the disease - recessive mutations in *IFNGR1* - was first described in 1996 (*172*, *173*) . A more recent example of Mendelian susceptibility to a single pathogen is chronic mucocutaneous candidiasis (CMC), which is the manifestation of unusual susceptibility to the commensal fungus *Candida albicans* and segregates as recessive or dominant due to mutations leading to impaired *IL-17* immunity (*174*) .

The notion of incomplete penetrance adds another layer of complexity to the understanding of host genetic susceptibility to infections. The monogenic architecture of susceptibility to one or a narrow range of infections can indeed be obscured when clinical symptoms are not present in all carriers of the mutant alleles. A prime example is herpes simplex encephalitis (HSE), the most common form of sporadic viral encephalitis in the western world, which is observed during primary infection with herpes simplex virus-1 (HSV-1). HSE is known since 1941, and has been rarely seen in familial form. Children who develop HSE upon HSV-1 infection are not particularly susceptible to other infections, and children with other PIDs are not more susceptible to HSE (*8*) , thus a genetic basis for the disease was not suspected. The first genetic etiology of HSE, recessive mutations in *UNC93B1*, was described in 2006 (*175*) . Since then, six other genetic causes of HSE have been described (*110*, *167*, *176*–*179*) : autosomal dominant and autosomal recessive *TLR3* deficiency, autosomal dominant *TBK1* deficiency, autosomal dominant *TRAF3* deficiency, autosomal dominant and autosomal recessive *TRIF* deficiency, and autosomal dominant *IRF3* deficiency. Because the disease is not systematically observed in individuals carrying the risk genotypes, susceptibility to HSE can be described as an example of monogenic but not fully penetrant susceptibility to infection.

Based on these examples and on similar findings that could not be explained using the previously accepted model, Alcaïs *et al.* proposed a more comprehensive model for the genetic architecture of infectious diseases in 2010 (Figure 3) (*180*) . In this model, age and fatality rate are considered important co-factors in deciphering the genetic architecture of infectious diseases. The monogenic component of the spectrum provides an explanation for the occurrence of life-threatening primary infections, which happen mostly in childhood, whereas polygenic/complex predisposition is more likely to explain secondary infections and infection reactivations, which occur mostly in older patients. Additional support for this model comes from a recent study by Brodin *et al.*, who investigated 204 immunological parameters in 210 healthy twins and showed that many of them become more variable with age, suggesting that the cumulative influence of environmental exposure alters the role of human genetics in susceptibility to infectious diseases in older patients (*181*) .

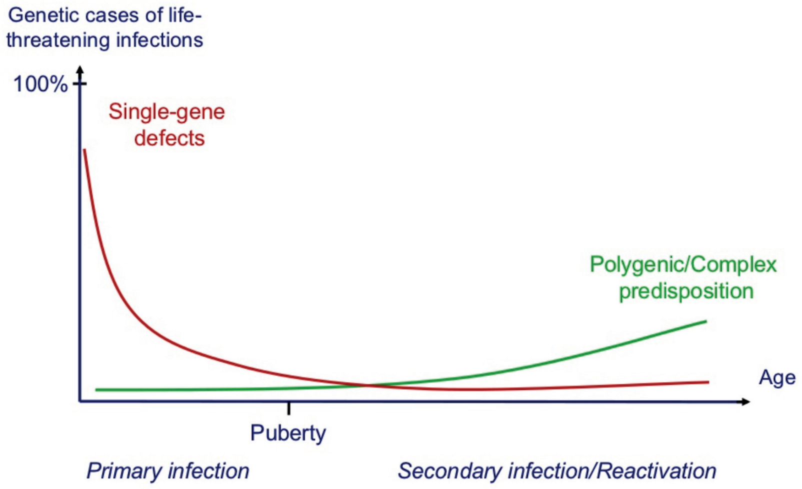

Figure 1.2: A model for genetic architecture of infectious diseases

Adopted from reference 180; Monogenic variants play an essential role in determining the outcome of primary infections. In contrast, secondary infections or reactivations of infections, which are more common among adults, are more likely to be due to polygenic predisposition (*180*).

1.4 Role of pathogen

Any infectious disease is the result of the interplay between host, pathogen and the environment. The evolutionary arm race between host and pathogen is a great force that shapes both host and pathogen genomes (*182*) . As discussed above, the clinical outcome of infections is influenced by variation in the human genome. On the other hand, genetic variation in pathogen genomes can also affect disease severity and clinical outcome. The power of pathogen genomics studies to transform our understanding of infectious disease became clear with the full genome sequencing of the first pathogen, *Haemophilus influenzae*, in 1995 (*183*) . Since then, pathogen genomics studies have been used to study drug resistance, transmission patterns, analysis of disease outbreaks, clinical diagnosis and pathogen virulence (*184*) . In a controversial study in 2012, Burke *et al.* used site-directed mutagenesis to mutate influenza A (H5N1) virus and showed that the virus can acquire mutations that greatly increase pathogenicity by enabling airborne transmission of the virus between mammals (*185*) . Multiple studies of influenza A (H1N1) virus reported association between mutations in haemagglutinin (HA), non-structural protein 1 (NS1) and polymerase basic protein 2 (PB2) and viral virulence (*186*) . Mutations in genotypes C and D of hepatitis B virus (HBV) have been shown to be associated with liver cirrhosis and progression to hepatocellular carcinoma, which are the most severe long-term complications of HBV infection (*187*) . These findings highlight the importance of including pathogen genomics in the study of human infections. Yet, most genomic studies of infectious diseases are conducted from the perspective of the host because the human genome is relatively conserved from generation to generation. Conversely, most pathogen genomes exhibit relatively high variability, making it challenging to discern mutations that have a modulating impact on host interaction and virulence from other neutral mutations (*188*) . With widespread access to HTS technologies, it is now possible to sequence viruses, bacteria, fungi or parasites isolated from each infected host and to look for pathogen genetic factors that can affect virulence. The result of such studies will help to get a more comprehensive picture of infectious disease pathogenesis and will pave the way for development of new preventive and therapeutic measures.

1.5 Justification of research

Infectious diseases are among the leading causes of human morbidity and mortality and the burden of infectious disorders is the greatest in the pediatric population. As described above, recent findings support the idea of a monogenic but non-Mendelian basis for many infectious diseases of childhood, and incomplete penetrance seems to be the rule rather than the exception. One strategy to increase the efficiency of gene discovery is to sequence individuals with the most dramatic clinical presentation, i.e. at the extreme of the phenotype distribution. The promise of this approach comes from the fact that variants of large effect sizes are under strong purifying selection and thus are found at low frequencies in the general population. Accordingly, for each trait with a normal distribution of phenotypic outcome in a population, individuals that are part of the tail of the distribution are more likely to carry variants of profound effect. Therefore, it is possible to design relatively small sequencing studies by focusing on individuals with extreme phenotypes, who are most likely to be informative genetically. Life-threatening susceptibility to individual pathogens in pediatric population is a rare, if not extremely rare, condition and extreme phenotype sampling is particularly beneficial to find the genetic etiology of these disorders.

In the next two chapters of this work, I present two host genomic studies, in which we used exome sequencing, bioinformatic analysis and functional follow-up to explore the genetic basis of extreme susceptibility to common infections in the pediatric population. The first is severe sepsis due to *Pseudomonas aeruginosa* (*P. aeruginosa*), an opportunistic bacterium that is found in abundance in soil, water and air, and rarely causes clinical symptoms in healthy individuals. However, in rare cases, it can cause bacteraemia and sepsis in previously healthy individuals, resulting in a high mortality rate. The second is severe lower respiratory tract infection due to respiratory syncytial virus (RSV). RSV is the most common viral respiratory infection in children, and usually causes only mild and self-limiting symptoms in healthy individuals. In the fourth chapter, I describe a viral genomic study, in which we used high-throughput sequencing to compare the genetic variability of two RSV strains extracted from nasopharyngeal aspirates (NPAs) of patients with extreme susceptibility to RSV with genomic variation of RSV strains extracted from patients with milder symptoms.

The global aim of this research is to identify human and pathogen genetic factors that contribute to inter-individual differences in susceptibility to *P. aeruginosa* and RSV. An improved understanding of the molecular basis of these differences could provide invaluable insight into host-pathogen interactions and disease pathogenesis.

1.6 References

1. D. G. Smith, The genetic hypothesis for susceptibility to lepromatous leprosy., *Human genetics* **50**, 163–77 (1979).
2. TM Daniel, The history of tuberculosis, *Respiratory medicine* (2006), doi:10.1016/j.rmed.2006.08.006.
3. J.-L. L. Casanova, L. Abel, The genetic theory of infectious diseases: a brief history and selected illustrations., *Annu Rev Genomics Hum Genet* **14**, 215–43 (2013).
4. A. Chadli, [Charles Nicolle and the accomplishments of his scientific thought]., *Arch Inst Pasteur Tunis* **63**, 3–14 (1986).
5. S. Kaufmann, Envisioning future strategies for vaccination against tuberculosis, *Nature Reviews Immunology* (2006), doi:10.1038/nri1920.
6. Cooke, Hill, Genetics of susceptibility to human infectious disease., *Nature reviews. Genetics* **2** (2001), doi:10.1038/35103577.
7. S. J. Chapman, A. V. Hill, Human genetic susceptibility to infectious disease., *Nat. Rev. Genet.* **13**, 175–88 (2012).
8. A. Alcaïs, L. Abel, J.-L. Casanova, Human genetics of infectious diseases: between proof of principle and paradigm, *Journal of Clinical Investigation* **119**, 2506–14 (2009).
9. Comstock, Tuberculosis in twins: a re-analysis of the Prophit survey., *The American review of respiratory disease* **117**, 621–4 (1978).
10. Sørensen, Nielsen, Andersen, Teasdale, Genetic and environmental influences on premature death in adult adoptees., *The New England journal of medicine* **318**, 727–32 (1988).
11. O. C. BRUTON, L. APT, D. GITLIN, C. A. JANEWAY, Absence of serum gamma globulins., *AMA Am J Dis Child* **84**, 632–6 (1952).
12. O. C. BRUTON, Agammaglobulinemia., *Pediatrics* **9**, 722–8 (1952).
13. A. V. Hill, Aspects of genetic susceptibility to human infectious diseases., *Annu. Rev. Genet.* **40**, 469–86 (2006).
14. A. Hill, The genomics and genetics of human infectious disease susceptibility, (2001).
15. A. V. Hill *et al.*, Common west African HLA antigens are associated with protection from severe malaria., *Nature* **352**, 595–600 (1991).
16. M. Thursz, R. Yallop, R. Goldin, C. Trepo, H. C. Thomas, Influence of MHC class II genotype on outcome of infection with hepatitis C virus. The HENCORE group. Hepatitis C European Network for Cooperative Research., *Lancet (London, England)* **354**, 2119–24 (1999).
17. M. Carrington *et al.*, HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage., *Science (New York, N.Y.)* **283**, 1748–52 (1999).
18. R. A. Kaslow *et al.*, Influence of combinations of human major histocompatibility complex genes on the course of HIV-1 infection., *Nature medicine* **2**, 405–11 (1996).
19. R. J. Klein *et al.*, Complement factor H polymorphism in age-related macular degeneration., *Science (New York, N.Y.)* **308**, 385–9 (2005).
20. J. Fellay *et al.*, A whole-genome association study of major determinants for host control of HIV-1., *Science (New York, N.Y.)* **317**, 944–7 (2007).
21. F. Pereyra *et al.*, The major genetic determinants of HIV-1 control affect HLA class I peptide presentation., *Science* **330**, 1551–7 (2010).

22. D. Ge *et al.*, Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance., *Nature* **461**, 399–401 (2009).

23. F.-R. R. Zhang *et al.*, Genomewide association study of leprosy., *N. Engl. J. Med.* **361**, 2609–18 (2009).

24. C. B. Moore *et al.*, Phenome-wide Association Study Relating Pretreatment Laboratory Parameters With Human Genetic Variants in AIDS Clinical Trials Group Protocols., *Open Forum Infect Dis* **2**, ofu113 (2015).

25. D. S. Lehmann *et al.*, Genome-wide association study of virologic response with efavirenz-containing or abacavir-containing regimens in AIDS clinical trials group protocols., *Pharmacogenet. Genomics* **25**, 51–9 (2015).

26. P. D. Leger *et al.*, Genome-wide association study of peripheral neuropathy with D-drug-containing regimens in AIDS Clinical Trials Group protocol 384., *J. Neurovirol.* **20**, 304–8 (2014).

27. P. J. McLaren *et al.*, Association study of common genetic variants and HIV-1 acquisition in 6,300 infected cases and 7,200 controls., *PLoS Pathog.* **9**, e1003515 (2013).

28. A. J. Levine *et al.*, Genome-wide association study of neurocognitive impairment and dementia in HIV-infected adults., *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **159B**, 669–83 (2012).

29. J. R. Lingappa *et al.*, Genomewide association study for determinants of HIV-1 acquisition and viral set point in HIV-1 serodiscordant couples with quantified virus exposure., *PLoS ONE* **6**, e28632 (2011).

30. M. R. Irvin *et al.*, Genes linked to energy metabolism and immunoregulatory mechanisms are associated with subcutaneous adipose tissue distribution in HIV-infected men., *Pharmacogenet. Genomics* **21**, 798–807 (2011).

31. S. Chantarangsu *et al.*, Genome-wide association study identifies variations in 6p21.3 associated with nevirapine-induced rash., *Clin. Infect. Dis.* **53**, 341–8 (2011).

32. V. Ramsuran *et al.*, Duffy-null-associated low neutrophil counts influence HIV-1 susceptibility in high-risk South African black women., *Clin. Infect. Dis.* **52**, 1248–56 (2011).

33. J. L. Troyer *et al.*, Genome-wide association study implicates PARD3B-based AIDS restriction., *J. Infect. Dis.* **203**, 1491–502 (2011).

34. S. M. Bol *et al.*, Genome-wide association study identifies single nucleotide polymorphism in DYRK1A associated with replication of HIV-1 in monocyte-derived macrophages., *PLoS ONE* **6**, e17190 (2011).

35. S. Petrovski *et al.*, Common human genetic variants and HIV-1 susceptibility: a genome-wide survey in a homogeneous African population., *AIDS* **25**, 513–8 (2011).

36. K. Pelak *et al.*, Host determinants of HIV-1 control in African Americans., *J. Infect. Dis.* **201**, 1141–9 (2010).

37. J. T. Herbeck *et al.*, Multistage genomewide association study identifies a locus at 1q41 associated with rate of HIV-1 disease progression to clinical AIDS., *J. Infect. Dis.* **201**, 618–26 (2010).

38. M. A. Ferreira *et al.*, Quantitative trait loci for CD4:CD8 lymphocyte ratio are associated with risk of type 1 diabetes and HIV-1 immune control., *Am. J. Hum. Genet.* **86**, 88–92 (2010).

39. J. Fellay *et al.*, Common genetic variation and the control of HIV-1 in humans., *PLoS Genet.* **5**, e1000791 (2009).

40. S. Shrestha *et al.*, A genome-wide association study of carotid atherosclerosis in HIV-infected men., *AIDS* **24**, 583–92 (2010).

41. S. Le Clerc *et al.*, Genomewide association study of a rapid progression cohort identifies new susceptibility alleles for AIDS (ANRS Genomewide Association Study 03)., *J. Infect. Dis.* **200**, 1194–201 (2009).

42. S. Limou *et al.*, Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02)., *J. Infect. Dis.* **199**, 419–26 (2009).

43. R. Rubicz *et al.*, A genome-wide integrative genomic study localizes genetic factors influencing antibodies against Epstein-Barr virus nuclear antigen 1 (EBNA-1)., *PLoS Genet.* **9**, e1003147 (2013).

44. D. Chen *et al.*, Genome-wide association study of HPV seropositivity., *Hum. Mol. Genet.* **20**, 4714–23 (2011).

45. X. Deng *et al.*, Genome wide association study (GWAS) of Chagas cardiomyopathy in Trypanosoma cruzi seropositive subjects., *PLoS ONE* **8**, e79629 (2013).

46. D. Miki *et al.*, Variation in the DEPDC5 locus is associated with progression to hepatocellular carcinoma in chronic hepatitis C virus carriers., *Nat. Genet.* **43**, 797–800 (2011).

47. A. L. Zignego *et al.*, Genome-wide association study of hepatitis C virus- and cryoglobulin-related vasculitis., *Genes Immun.* **15**, 500–5 (2014).

48. D. Miki *et al.*, HLA-DQB1*03 confers susceptibility to chronic hepatitis C in Japanese: a genome-wide association study., *PLoS ONE* **8**, e84226 (2013).

49. Y. Tanaka *et al.*, Genome-wide association study identified ITPA/DDRGK1 variants reflecting thrombocytopenia in pegylated interferon and ribavirin therapy for chronic hepatitis C., *Hum. Mol. Genet.* **20**, 3507–16 (2011).

50. H. Ochi *et al.*, IL-28B predicts response to chronic hepatitis C therapy--fine-mapping and replication study in Asian populations., *J. Gen. Virol.* **92**, 1071–81 (2011).

51. J. Fellay *et al.*, ITPA gene variants protect against anaemia in patients treated for chronic hepatitis C., *Nature* **464**, 405–8 (2010).

52. A. Rauch *et al.*, Genetic variation in IL28B is associated with chronic hepatitis C and treatment failure: a genome-wide association study., *Gastroenterology* **138**, 1338–45, 1345.e1–7 (2010).

53. V. Suppiah *et al.*, IL28B is associated with response to chronic hepatitis C interferon-alpha and ribavirin therapy., *Nat. Genet.* **41**, 1100–4 (2009).

54. A. J. Thompson *et al.*, Genome-wide association study of interferon-related cytopenia in chronic hepatitis C patients., *J. Hepatol.* **56**, 313–9 (2012).

55. P. Duggal *et al.*, Genome-wide association study of spontaneous resolution of hepatitis C virus infection: data from multiple cohorts., *Ann. Intern. Med.* **158**, 235–45 (2013).

56. P. J. Clark *et al.*, Interleukin 28B polymorphisms are the only common genetic variants associated with low-density lipoprotein cholesterol (LDL-C) in genotype-1 chronic hepatitis C and determine the association between LDL-C and treatment response., *J. Viral Hepat.* **19**, 332–40 (2012).

57. C. M. Lange *et al.*, Serum ferritin levels are associated with a distinct phenotype of chronic hepatitis C poorly responding to pegylated interferon-alpha and ribavirin therapy., *Hepatology* **55**, 1038–47 (2012).

58. E. Patin *et al.*, Genome-wide association study identifies variants associated with progression of liver fibrosis from HCV infection., *Gastroenterology* **143**, 1244–52.e1–12 (2012).

59. Y. Urabe *et al.*, A genome-wide association study of HCV-induced liver cirrhosis in the Japanese population identifies novel susceptibility loci at the MHC region., *J. Hepatol.* **58**, 875–82 (2013).

60. D. Nelson *et al.*, Genome-wide association study to characterize serum bilirubin elevations in patients with HCV treated with GS-9256, an HCV NS3 serine protease inhibitor., *Antivir. Ther. (Lond.)* **19**, 679–86 (2014).

61. E. R. Chimusa *et al.*, Genome-wide association study of ancestry-specific TB risk in the South African Coloured population., *Hum. Mol. Genet.* **23**, 796–809 (2014).

62. T. Thye *et al.*, Common variants at 11p13 are associated with susceptibility to tuberculosis., *Nat. Genet.* **44**, 257–9 (2012).

63. E. Png *et al.*, A genome wide association study of pulmonary tuberculosis susceptibility in Indonesians., *BMC Med. Genet.* **13**, 5 (2012).

64. J. Curtis *et al.*, Susceptibility to tuberculosis is associated with variants in the ASAP1 gene encoding a regulator of dendritic cell migration., *Nat. Genet.* **47**, 523–7 (2015).

65. T. Thye *et al.*, Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2., *Nat. Genet.* **42**, 739–41 (2010).

66. M. Jallow *et al.*, Genome-wide and fine-resolution association analysis of malaria in West Africa., *Nature genetics* **41**, 657–65 (2009).

67. G. Band *et al.*, Imputation-based meta-analysis of severe malaria in three African populations., *PLoS Genet.* **9**, e1003509 (2013).

68. C. Timmann *et al.*, Genome-wide association study indicates two novel resistance loci for severe malaria., *Nature* **489**, 443–6 (2012).

69. K. Ding *et al.*, Genetic variants that confer resistance to malaria are associated with red blood cell traits in African-Americans: an electronic medical record-based genome-wide association study., *G3 (Bethesda)* **3**, 1061–8 (2013).

70. Y. Kamatani *et al.*, A genome-wide association study identifies variants in the HLA-DP locus associated with chronic hepatitis B in Asians., *Nat. Genet.* **41**, 591–5 (2009).

71. H. Mbarek *et al.*, A genome-wide association study of chronic hepatitis B identified novel risk locus in a Japanese population., *Hum. Mol. Genet.* **20**, 3884–92 (2011).

72. E. Png *et al.*, A genome-wide association study of hepatitis B vaccine response in an Indonesian population reveals multiple independent risk variants in the HLA region., *Hum. Mol. Genet.* **20**, 3893–8 (2011).

73. L. Liu *et al.*, A genome-wide association study with DNA pooling identifies the variant rs11866328 in the GRIN2A gene that affects disease progression of chronic HBV infection., *Viral Immunol.* **24**, 397–402 (2011).

74. N. Nishida *et al.*, Genome-wide association study confirming association of HLA-DP with protection against chronic hepatitis B and viral clearance in Japanese and Korean., *PLoS ONE* **7**, e39175 (2012).

75. Y. J. Kim *et al.*, A genome-wide association study identified new variants associated with the risk of chronic hepatitis B., *Hum. Mol. Genet.* **22**, 4233–8 (2013).

76. Z. Hu *et al.*, New loci associated with chronic hepatitis B virus infection in Han Chinese., *Nat. Genet.* **45**, 1499–503 (2013).

77. L. Pan *et al.*, A genome-wide association study identifies polymorphisms in the HLA-DR region associated with non-response to hepatitis B vaccination in Chinese Han populations., *Hum. Mol. Genet.* **23**, 2210–9 (2014).
78. S.-W. W. Chang *et al.*, A genome-wide association study on chronic HBV infection and its clinical progression in male Han-Taiwanese., *PLoS ONE* **9**, e99724 (2014).
79. D.-K. K. Jiang *et al.*, Genetic variants in five novel loci including CFB and CD40 predispose to chronic hepatitis B., *Hepatology* **62**, 118–28 (2015).
80. J. Zhou *et al.*, A functional variation in CD55 increases the severity of 2009 pandemic H1N1 influenza A virus infection., *J. Infect. Dis.* **206**, 495–503 (2012).
81. E. K. Miller *et al.*, Atopy history and the genomics of wheezing after influenza vaccination in children 6-59 months of age., *Vaccine* **29**, 3431–7 (2011).
82. A. Rautanen *et al.*, Genome-wide association study of survival from sepsis due to pneumonia: an observational cohort study., *Lancet Respir Med* **3**, 53–60 (2015).
83. Z. Ye *et al.*, Genome wide association study of SNP-, gene-, and pathway-based approaches to identify genes influencing susceptibility to Staphylococcus aureus infections., *Front Genet* **5**, 125 (2014).
84. S. J. Dunstan *et al.*, Variation at HLA-DRB1 is associated with resistance to enteric fever., *Nat. Genet.* **46**, 1333–6 (2014).
85. H. Liu *et al.*, Discovery of six new susceptibility loci and analysis of pleiotropic effects in leprosy., *Nat. Genet.* **47**, 267–71 (2015).
86. F. Zhang *et al.*, Identification of two new loci at IL23R and RAB32 that influence susceptibility to leprosy., *Nat. Genet.* **43**, 1247–51 (2011).
87. M. Fakiola *et al.*, Common variants in the HLA-DRB1-HLA-DQA1 HLA class II region are associated with susceptibility to visceral leishmaniasis., *Nat. Genet.* **45**, 208–13 (2013).
88. S. Davila *et al.*, Genome-wide association study identifies variants in the CFH region associated with host susceptibility to meningococcal disease., *Nature genetics* **42**, 772–6 (2010).
89. D. R. Crosslin *et al.*, Genetic variation in the HLA region is associated with susceptibility to herpes zoster., *Genes Immun.* **16**, 1–7 (2015).
90. C. C. Khor *et al.*, Genome-wide association study identifies susceptibility loci for dengue shock syndrome at MICB and PLCE1., *Nat. Genet.* **43**, 1139–41 (2011).
91. P. Sanchez-Juan *et al.*, Genome-wide study links MTMR7 gene to variant Creutzfeldt-Jakob risk., *Neurobiol. Aging* **33**, 1487.e21–8 (2012).
92. S. Mead *et al.*, Genetic risk factors for variant Creutzfeldt-Jakob disease: a genome-wide association study., *Lancet Neurol* **8**, 57–66 (2009).
93. W. Bush, J. Moore, Chapter 11: Genome-wide association studies., *PLoS computational biology* **8**, e1002822 (2012).
94. CS Pareek, R Smoczynski, A Tretyn, Sequencing technologies and genome sequencing, *Journal of applied genetics* (2011), doi:10.1007/s13353-011-0057-x.
95. M. Byun *et al.*, Whole-exome sequencing-based discovery of STIM1 deficiency in a child with fatal classic Kaposi sarcoma., *J. Exp. Med.* **207**, 2307–12 (2010).
96. A. Bolze *et al.*, Whole-exome-sequencing-based discovery of human FADD deficiency., *Am. J. Hum. Genet.* **87**, 873–81 (2010).
97. F.-Y. Y. Li *et al.*, Second messenger role for Mg2+ revealed by human T-cell immunodeficiency., *Nature* **475**, 471–6 (2011).

98. J. C. de Greef *et al.*, Mutations in ZBTB24 are associated with immunodeficiency, centromeric instability, and facial anomalies syndrome type 2., *Am. J. Hum. Genet.* **88**, 796–804 (2011).

99. L. Liu *et al.*, Gain-of-function human STAT1 mutations impair IL-17 immunity and underlie chronic mucocutaneous candidiasis., *J. Exp. Med.* **208**, 1635–48 (2011).

100. F. L. van de Veerdonk *et al.*, STAT1 mutations in autosomal dominant chronic mucocutaneous candidiasis., *N. Engl. J. Med.* **365**, 54–61 (2011).

101. C. N. Hahn *et al.*, Heritable GATA2 mutations associated with familial myelodysplastic syndrome and acute myeloid leukemia., *Nat. Genet.* **43**, 1012–7 (2011).

102. P. Ostergaard *et al.*, Mutations in GATA2 cause primary lymphedema associated with a predisposition to acute myeloid leukemia (Emberger syndrome)., *Nat. Genet.* **43**, 929–31 (2011).

103. R. E. Dickinson *et al.*, Exome sequencing identifies GATA-2 mutation as the cause of dendritic cell, monocyte, B and NK lymphoid deficiency., *Blood* **118**, 2656–8 (2011).

104. A. P. Hsu *et al.*, Mutations in GATA2 are associated with the autosomal dominant and sporadic monocytopenia and mycobacterial infection (MonoMAC) syndrome., *Blood* **118**, 2653–5 (2011).

105. A. Crequer *et al.*, Human RHOH deficiency causes T cell defects and susceptibility to EV-HPV infections., *J. Clin. Invest.* **122**, 3239–47 (2012).

106. A. L. Snow *et al.*, Congenital B cell lymphocytosis explained by novel germline CARD11 mutations., *J. Exp. Med.* **209**, 2247–61 (2012).

107. L. Gineau *et al.*, Partial MCM4 deficiency in patients with growth retardation, adrenal insufficiency, and natural killer cell deficiency., *J. Clin. Invest.* **122**, 821–32 (2012).

108. C. R. Hughes *et al.*, MCM4 mutation causes adrenal failure, short stature, and natural killer cell deficiency in humans., *J. Clin. Invest.* **122**, 814–20 (2012).

109. J. Pachlopnik Schmid *et al.*, Polymerase ε1 mutation in a human syndrome with facial dysmorphism, immunodeficiency, livedo, and short stature ("FILS syndrome")., *J. Exp. Med.* **209**, 2323–30 (2012).

110. M. Herman *et al.*, Heterozygous TBK1 mutations impair TLR3 immunity and underlie herpes simplex encephalitis of childhood., *J. Exp. Med.* **209**, 1567–82 (2012).

111. C. T. Jordan *et al.*, Rare and common variants in CARD14, encoding an epidermal regulator of NF-kappaB, in psoriasis., *Am. J. Hum. Genet.* **90**, 796–808 (2012).

112. C. T. Jordan *et al.*, PSORS2 is due to mutations in CARD14., *Am. J. Hum. Genet.* **90**, 784–95 (2012).

113. D. Fuchs-Telem *et al.*, Familial pityriasis rubra pilaris is caused by mutations in CARD14., *Am. J. Hum. Genet.* **91**, 163–70 (2012).

114. M. J. Ombrello *et al.*, Cold urticaria, immunodeficiency, and autoimmunity related to PLCG2 deletions., *N. Engl. J. Med.* **366**, 330–8 (2012).

115. Q. Zhou *et al.*, A hypermorphic missense mutation in PLCG2, encoding phospholipase Cγ2, causes a dominantly inherited autoinflammatory disease with immunodeficiency., *Am. J. Hum. Genet.* **91**, 713–20 (2012).

116. M. E. Conley *et al.*, Agammaglobulinemia and absent B lineage cells in a patient lacking the p85α subunit of PI3K., *J. Exp. Med.* **209**, 463–70 (2012).

117. G. I. Rice *et al.*, Mutations in ADAR1 cause Aicardi-Goutières syndrome associated with a type I interferon signature., *Nat. Genet.* **44**, 1243–8 (2012).

118. D. Bogunovic *et al.*, Mycobacterial disease and impaired IFN-γ immunity in humans with inherited ISG15 deficiency., *Science* **337**, 1684–8 (2012).

119. B. Boisson *et al.*, Immunodeficiency, autoinflammation and amylopectinosis in humans with inherited HOIL-1 and LUBAC deficiency., *Nat. Immunol.* **13**, 1178–86 (2012).

120. J. M. van Montfrans *et al.*, CD27 deficiency is associated with combined immunodeficiency and persistent symptomatic EBV viremia., *J. Allergy Clin. Immunol.* **129**, 787–793.e6 (2012).

121. A. Alangari *et al.*, LPS-responsive beige-like anchor (LRBA) gene mutation in a family with inflammatory bowel disease and combined immunodeficiency., *J. Allergy Clin. Immunol.* **130**, 481–8.e2 (2012).

122. S. O. Burns *et al.*, LRBA gene deletion in a patient presenting with autoimmunity without hypogammaglobulinemia., *J. Allergy Clin. Immunol.* **130**, 1428–32 (2012).

123. G. Lopez-Herrera *et al.*, Deleterious mutations in LRBA are associated with a syndrome of immune deficiency and autoimmunity., *Am. J. Hum. Genet.* **90**, 986–1001 (2012).

124. H. H. Jabara *et al.*, A homozygous mucosa-associated lymphoid tissue 1 (MALT1) mutation in a family with combined immunodeficiency., *J. Allergy Clin. Immunol.* **132**, 151–8 (2013).

125. D. Kotlarz *et al.*, Loss-of-function mutations in the IL-21 receptor gene cause a primary immunodeficiency syndrome., *J. Exp. Med.* **210**, 433–43 (2013).

126. P. Stepensky *et al.*, Deficiency of caspase recruitment domain family, member 11 (CARD11), causes profound combined immunodeficiency in human subjects., *J. Allergy Clin. Immunol.* **131**, 477–85.e1 (2013).

127. J. Greil *et al.*, Whole-exome sequencing links caspase recruitment domain 11 (CARD11) inactivation to severe combined immunodeficiency., *J. Allergy Clin. Immunol.* **131**, 1376–83.e3 (2013).

128. M. Byun *et al.*, Inherited human OX40 deficiency underlying classic Kaposi sarcoma of childhood., *J. Exp. Med.* **210**, 1743–59 (2013).

129. M. E. Samuels *et al.*, Exome sequencing identifies mutations in the gene TTC7A in French-Canadian cases with hereditary multiple intestinal atresia., *J. Med. Genet.* **50**, 324–9 (2013).

130. R. Chen *et al.*, Whole-exome sequencing identifies tetratricopeptide repeat domain 7A (TTC7A) mutations for combined immunodeficiency with intestinal atresias., *J. Allergy Clin. Immunol.* **132**, 656–664.e17 (2013).

131. B. Boisson *et al.*, An ACT1 mutation selectively abolishes interleukin-17 responses in humans with chronic mucocutaneous candidiasis., *Immunity* **39**, 676–86 (2013).

132. B. Boisson *et al.*, A recurrent dominant negative E47 mutation causes agammaglobulinemia and BCR(-) B cells., *J. Clin. Invest.* **123**, 4781–5 (2013).

133. H.-Y. Y. Wang *et al.*, Antibody deficiency associated with an inherited autosomal dominant mutation in TWEAK., *Proc. Natl. Acad. Sci. U.S.A.* **110**, 5127–32 (2013).

134. K. Chen *et al.*, Germline mutations in NFKB2 implicate the noncanonical NF-κB pathway in the pathogenesis of common variable immunodeficiency., *Am. J. Hum. Genet.* **93**, 812–24 (2013).

135. H. S. Kuehn *et al.*, Loss-of-function of the protein kinase C δ (PKCδ) causes a B-cell lymphoproliferative syndrome in humans., *Blood* **121**, 3117–25 (2013).

136. E. Salzer *et al.*, B-cell deficiency and severe autoimmunity caused by deficiency of protein kinase C δ., *Blood* **121**, 3112–6 (2013).

137. A. Belot *et al.*, Protein kinase cδ deficiency causes mendelian systemic lupus erythematosus with B cell-defective apoptosis and hyperproliferation., *Arthritis Rheum.* **65**, 2161–71 (2013).

138. I. Angulo *et al.*, Phosphoinositide 3-kinase δ gene mutation predisposes to respiratory infection and airway damage., *Science* **342**, 866–71 (2013).

139. A. Bolze *et al.*, Ribosomal protein SA haploinsufficiency in humans with isolated congenital asplenia., *Science* **340**, 976–8 (2013).

140. P. Stepensky *et al.*, The Thr224Asn mutation in the VPS45 gene is associated with the congenital neutropenia and primary myelofibrosis of infancy., *Blood* **121**, 5078–87 (2013).

141. T. Vilboux *et al.*, A congenital neutrophil defect syndrome associated with mutations in VPS45., *N. Engl. J. Med.* **369**, 54–65 (2013).

142. Z. Deng *et al.*, Inherited mutations in the helicase RTEL1 cause telomere dysfunction and Hoyeraal-Hreidarsson syndrome., *Proc. Natl. Acad. Sci. U.S.A.* **110**, E3408–16 (2013).

143. T. Le Guen *et al.*, Human RTEL1 deficiency causes Hoyeraal-Hreidarsson syndrome with short telomeres and genome instability., *Hum. Mol. Genet.* **22**, 3239–49 (2013).

144. U. Pannicke *et al.*, Deficiency of innate and acquired immunity caused by an IKBKB mutation., *N. Engl. J. Med.* **369**, 2504–14 (2013).

145. S. O. Burns *et al.*, Immunodeficiency and disseminated mycobacterial infection associated with homozygous nonsense mutation of IKKβ., *J. Allergy Clin. Immunol.* **134**, 215–8 (2014).

146. J. M. Torres *et al.*, Inherited BCL10 deficiency impairs hematopoietic and nonhematopoietic immunity., *J. Clin. Invest.* **124**, 5239–48 (2014).

147. E. Martin *et al.*, CTP synthase 1 deficiency in humans reveals its central role in lymphocyte proliferation., *Nature* **510**, 288–92 (2014).

148. P. K. Chakraborty *et al.*, Mutations in TRNT1 cause congenital sideroblastic anemia with immunodeficiency, fevers, and developmental delay (SIFD)., *Blood* **124**, 2867–71 (2014).

149. E. Salzer *et al.*, Early-onset inflammatory bowel disease and common variable immunodeficiency-like disease caused by IL-21 deficiency., *J. Allergy Clin. Immunol.* **133**, 1651–9.e12 (2014).

150. H. S. Kuehn *et al.*, Immune dysregulation in human subjects with heterozygous germline mutations in CTLA4., *Science* **345**, 1623–7 (2014).

151. N. Jeremiah *et al.*, Inherited STING-activating mutation underlies a familial inflammatory syndrome with lupus-like manifestations., *J. Clin. Invest.* **124**, 5516–20 (2014).

152. N. Romberg *et al.*, Mutation of NLRC4 causes a syndrome of enterocolitis and autoinflammation., *Nat. Genet.* **46**, 1135–9 (2014).

153. K. L. Willmann *et al.*, Biallelic loss-of-function mutation in NIK causes a primary immunodeficiency with multifaceted aberrant lymphoid immunity., *Nat Commun* **5**, 5360 (2014).

154. H. Kocak *et al.*, Hoyeraal-Hreidarsson syndrome caused by a germline mutation in the TEL patch of the telomere protein TPP1., *Genes Dev.* **28**, 2090–102 (2014).

155. K. Boztug *et al.*, JAGN1 deficiency causes aberrant myeloid cell homeostasis and congenital neutropenia., *Nat. Genet.* **46**, 1021–7 (2014).

156. S. Kracker *et al.*, An inherited immunoglobulin class-switch recombination deficiency associated with a defect in the INO80 chromatin remodeling complex., *J. Allergy Clin. Immunol.* **135**, 998–1007.e6 (2015).

157. M. Aavikko *et al.*, Whole-Genome Sequencing Identifies STAT4 as a Putative Susceptibility Gene in Classic Kaposi Sarcoma., *J. Infect. Dis.* **211**, 1842–51 (2015).

158. Q. Zhou *et al.*, Early-onset stroke and vasculopathy associated with mutations in ADA2., *N. Engl. J. Med.* **370**, 911–20 (2014).

159. P. Navon Elkan *et al.*, Mutant adenosine deaminase 2 in a polyarteritis nodosa vasculopathy., *N. Engl. J. Med.* **370**, 921–31 (2014).

160. G. I. Rice *et al.*, Gain-of-function mutations in IFIH1 cause a spectrum of human disease phenotypes associated with upregulated type I interferon signaling., *Nat. Genet.* **46**, 503–9 (2014).

161. M. J. Ciancanelli *et al.*, Infectious disease. Life-threatening influenza and impaired interferon amplification in human IRF7 deficiency., *Science* **348**, 448–53 (2015).

162. K. Dobbs *et al.*, Inherited DOCK2 Deficiency in Patients with Early-Onset Invasive Infections., *N. Engl. J. Med.* **372**, 2409–22 (2015).

163. B. Boisson *et al.*, Human HOIP and LUBAC deficiency underlies autoinflammation, immunodeficiency, amylopectinosis, and lymphangiectasia., *J. Exp. Med.* **212**, 939–51 (2015).

164. A. Sassi *et al.*, Hypomorphic homozygous mutations in phosphoglucomutase 3 (PGM3) impair immunity and increase serum IgE levels., *J. Allergy Clin. Immunol.* **133**, 1410–9, 1419.e1–13 (2014).

165. J. Record *et al.*, Immunodeficiency and severe susceptibility to bacterial infection associated with a loss-of-function homozygous mutation of MKL1., *Blood* **126**, 1527–35 (2015).

166. S. Okada *et al.*, IMMUNODEFICIENCIES. Impairment of immunity to Candida and Mycobacterium in humans with bi-allelic RORC mutations., *Science* **349**, 606–13 (2015).

167. L. L. Andersen *et al.*, Functional IRF3 deficiency in a patient with herpes simplex encephalitis., *J. Exp. Med.* **212**, 1371–9 (2015).

168. J.-L. L. Casanova, Severe infectious diseases of childhood as monogenic inborn errors of immunity., *Proc. Natl. Acad. Sci. U.S.A.* (2015), doi:10.1073/pnas.1521651112.

169. W. LUTZ, [Not Available]., *Dermatologica* **92**, 333–5 (1946).

170. N. Ramoz *et al.*, Mutations in two adjacent novel genes are associated with epidermodysplasia verruciformis., *Nat. Genet.* **32**, 579–81 (2002).

171. A. J. Coffey *et al.*, Host response to EBV infection in X-linked lymphoproliferative disease results from mutations in an SH2-domain encoding gene., *Nat. Genet.* **20**, 129–35 (1998).

172. M. J. Newport *et al.*, A mutation in the interferon-gamma-receptor gene and susceptibility to mycobacterial infection., *N. Engl. J. Med.* **335**, 1941–9 (1996).

173. E. Jouanguy *et al.*, Interferon-gamma-receptor deficiency in an infant with fatal bacille Calmette-Guérin infection., *N. Engl. J. Med.* **335**, 1956–61 (1996).

174. A. Puel *et al.*, Chronic mucocutaneous candidiasis in humans with inborn errors of interleukin-17 immunity., *Science* **332**, 65–8 (2011).

175. A. Casrouge *et al.*, Herpes simplex virus encephalitis in human UNC-93B deficiency., *Science* **314**, 308–12 (2006).

176. S.-Y. Y. Zhang *et al.*, TLR3 deficiency in patients with herpes simplex encephalitis., *Science* **317**, 1522–7 (2007).

177. Y. Guo *et al.*, Herpes simplex virus encephalitis in a patient with complete TLR3 deficiency: TLR3 is otherwise redundant in protective immunity., *J. Exp. Med.* **208**, 2083–98 (2011).

178. R. Pérez de Diego *et al.*, Human TRAF3 adaptor molecule deficiency leads to impaired Toll-like receptor 3 response and susceptibility to herpes simplex encephalitis., *Immunity* **33**, 400–11 (2010).

179. V. Sancho-Shimizu *et al.*, Herpes simplex encephalitis in children with autosomal recessive and dominant TRIF deficiency., *J. Clin. Invest.* **121**, 4889–902 (2011).

180. A. Alcaïs *et al.*, Life-threatening infectious diseases of childhood: single-gene inborn errors of immunity?, *Ann. N. Y. Acad. Sci.* **1214**, 18–33 (2010).

181. P Brodin, V Jojic, T Gao, S Bhattacharya, C. Angel, Variation in the human immune system is largely driven by non-heritable influences, *Cell* (2015).

182. M. Sironi, R. Cagliani, D. Forni, M. Clerici, Evolutionary insights into host-pathogen interactions from mammalian sequence data, *Nature Reviews Genetics* **16** (2015), doi:10.1038/nrg3905.

183. R. D. Fleischmann *et al.*, Whole-genome random sequencing and assembly of Haemophilus influenzae Rd., *Science (New York, N.Y.)* **269**, 496–512 (1995).

184. R. K. Bains, Human infectious diseases in the genomics era: where do we go from here?, *Genome Biol.* **15**, 529 (2014).

185. S. Herfst *et al.*, Airborne transmission of influenza A/H5N1 virus between ferrets., *Science (New York, N.Y.)* **336**, 1534–41 (2012).

186. E. A. Goka, P. J. Vallely, K. J. Mutton, P. E. Klapper, Mutations associated with severity of the pandemic influenza A(H1N1)pdm09 in humans: a systematic review and meta-analysis of epidemiological evidence., *Arch. Virol.* **159**, 3167–83 (2014).

187. M. Sunbul, Hepatitis B virus genotypes: global distribution and clinical importance., *World J. Gastroenterol.* **20**, 5427–34 (2014).

188. C. C. Khor, M. L. Hibberd, Host-pathogen interactions revealed by human genome-wide surveys., *Trends Genet.* **28**, 233–43 (2012).

Chapter 2 Exome sequencing reveals primary immunodeficiencies in children with *Pseudomonas aeruginosa* sepsis

2.1 Introduction

Pseudomonas aeruginosa (P. aeruginosa) is an aerobic gram-negative bacterium commonly found in soil, water and plants. This opportunistic pathogen causes invasive infections in immunosuppressed and hospitalized patients, and represents a major cause of healthcare-related infections. Sepsis due to community-acquired *P. aeruginosa* is extremely rare in apparently healthy children, and carries very high mortality (*1, 2*). Few case reports have described a range of primary immunodeficiencies (PID) in children with *P. aeruginosa* sepsis (*3*, *4*). While genetic alterations leading to PID are kept at very low frequency in the general population through negative selection, these variants may be found at higher frequency in children presenting with severe infections in the absence of known comorbidities (*5*). Here, we used exome sequencing to identify genetic variants causing PIDs in previously healthy children with *P. aeruginosa* sepsis.

2.2 Material and methods

2.2.1 Patients

Children below 60 month-old with community-acquired *P. aeruginosa* bloodstream sepsis (*6*) were eligible. Patients with existing comorbid conditions (prematurity, congenital malformations, previous surgery, immunosuppression, known immunodeficiency, chronic diseases) were excluded. The ethics committees of participating centers approved the study. Written informed consent from parents was obtained. Patients were recruited prospectively, and retrospectively in hospital databases and infectious diseases networks. Whenever possible, DNA was also collected from the parents.

2.2.2 DNA extraction and exome sequencing

Human genomic DNA was extracted from whole blood (N=8) or from fibroblasts cultured from frozen skin biopsy, frozen lymphatic tissue, or paraffin-fixed histology slides in one patient each. Exome sequencing libraries were prepared using Agilent SureSelect (V5 spanning 50.4 Mb). Cluster generation was performed using the Illumina TruSeq PE Cluster Kit v5 reagents. The resulting libraries were sequenced as 100 basepair long, paired-end reads on Illumina HiSeq 2500 using TruSeq SBS Kit v5 reagents.

2.2.3 Alignment of sequencing reads, variant calling and annotation

Raw sequencing reads were processed using CASAVA v1.82. Reads were aligned to human reference genome hg19 using BWA v0.6.2 (*7*). PCR duplicates were removed using Picard. Single nucleotide variants (SNVs) and small insertion and deletions (Indels) were called by the HaplotypeCaller in Genome Analysis Toolkit v3.1-1 using a multi-sample approach. We used SnpEff (*8*) to predict the functional impact of variants. The presence of a functionally relevant variant was confirmed using custom Taqman® genotyping assays.

2.2.4 Identification of potential causal variants

We restricted analyses to nonsynonymous exonic variants with a minor allele frequency (MAF) < 1%. For parent-child trios, we analysed each family separately assuming multiple inheritance models (autosomal recessive, autosomal dominant, X-linked recessive, compound heterozygous recessive). Only variants with at least 10X coverage in both parents and the offspring were included in the *de novo* analysis. For individual patients, we searched for rare, nonsynonymous variants assuming autosomal recessive, or X-linked recessive inheritance. We did not consider a mutation as potentially causal if it was present in the Exome Aggregation Consortium database (ExAC, http://exac.broadinstitute.org) or in a set of 533 in-house control exomes in the same zygosity form (i.e. heterozygous or homozygous) as in our patients. In addition, we ran a targeted search for rare, nonsynonymous variants in a list of 251 known PID genes (*9*).

2.2.5 Gene-annotation enrichment analysis

We used the Database for Annotation, Visualization and Integrated Discovery (*10*) (DAVID, https://david.ncifcrf.gov/home.jsp) v6.7 with default options and highest classification stringency for functional annotation clustering of PID genes carrying rare, nonsynonymous variants. We used hypergeometric test for assessing the significance of the gene enrichment result.

2.3 Results

2.3.1 Clinical presentation and outcomes

We identified 11 previously healthy children with community-acquired *P. aeruginosa* bacteremia from whom DNA could be obtained. The presenting age ranged from 6 months to 4 years, and 7/11 (64%) of patients were male. One surviving patient presented with acute abdomen, and *P. aeruginosa* bacteraemia was thought to result from intra-abdominal perforation. Another surviving patient was diagnosed with urosepsis and *P. aeruginosa* grew in urine and blood. Post-mortem examination reports were available in seven cases. In four, *P. aeruginosa* grew in high quantities from nasal and oral mucosal swabs, tracheal secretions, and lungs in one case patient presented with ecthyma gangraenosum (Table 2.1, Figure 2.1).

Table 2.1: Patients' demographic information
Demographic and clinical characteristics of included children with community-acquired *Pseudomonas aeruginosa* septicaemia.
CRP: C-reactive protein; C9def: Complement C9 deficiency; WCC: white cell count (*10^9/L)

Sample	DNA source	Phenotype	Ethnicity	Age (mo)	Sex	Consanguinity	Viral co-infection	WCC	CRP	Clinical focus	Previous history
P101	blood	fatal	caucasian	24	M	no	Parainfluenza 3	1.2	136	Pneumonia	recurrent middle ear infections
P201	blood	fatal	caucasian	13	M	no	HHV-6	0.6	NA	Septic shock	recurrent middle ear infections
P301	blood	fatal	caucasian	9	F	no	Parainfluenza 3	1.3	46	Ecthyma gangraenosum	nil
P401	blood	survived	south american	8	M	no	none	2	132	Acute abdomen	nil
P501	blood	fatal	asian	26	M	yes	Varicella	2.4	36	Septic shock	recurrent respiratory infections
P601	lymph node	fatal	african	31	F	no	none	1	NA	Pneumonia	nil
P801	blood	survived	caucasian	56	M	no	none	26.9	235	Urosepsis	nil
P701	blood	survived	caucasian	7	M	no	none	1.2	301	Septic shock	recurrent infection-associated neutropenia
P901	fibroblasts	fatal	asian	9	M	no	hMPV, RSV	0.8	NA	Septic shock	nil
P1001	paraffin slides	fatal	caucasian	30	M	no	none	1.2	213	Pneumonia	recurrent respiratory infections
P1101	blood	fatal	asian	26	M	no	none	1.6	157	Septic shock	recurrent respiratory infections

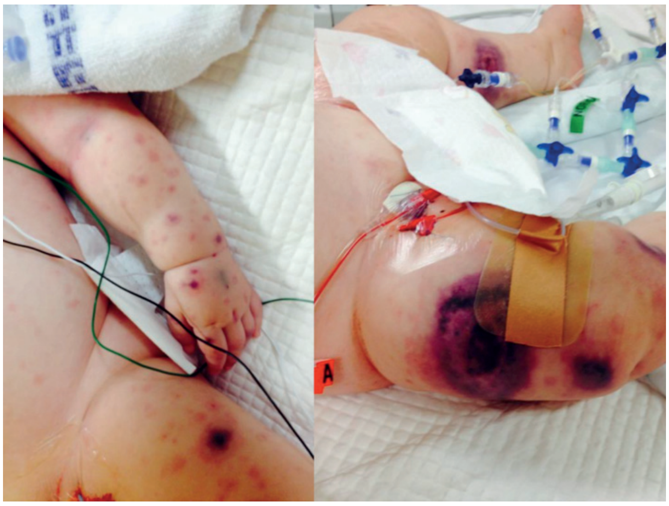

Figure 2.1: Ecthyma grangrensum

Ecthyma grangrenosum is a highly suggestive rapidly progressive purpuric skin lesion seen in a minority of children with *P. aeruginosa* bacteremia, is shown in one of the study patients (with permission from parents).

2.3.2 Exome sequencing

The mean coverage was 66x and 96% of target bases achieved at least 10x coverage. 115,604 SNVs and 10,814 indels passed GATK quality control: 28,795 synonymous variants, 28,284 nonsynonymous variants, 713 in-frame indels, 660 frame-shift indels, 283 stop-gained and 151 splice-site variants (Tables 2.2-4.4).

Table 2.2: BWA alignment metrics

PF: pass filter

Sample	PF unique reads	PF unique reads aligned	Off-bait (%)	Mean bait coverage	At least 2X coverage (%)	At least 30X coverage (%)
P101	97.25	96.66	0.23	65.71	99.39	80.6249
P102	96.90	96.54	0.24	59.08	99.26	77.17
P103	96.39	95.99	0.26	62.26	99.39	78.68
P201	97.00	94.46	0.25	50.43	99.15	67.86
P202	98.94	96.55	0.25	17.97	97.55	15.88
P203	97.18	96.52	0.25	72.83	99.37	82.47
P301	97.77	96.40	0.25	58.37	99.24	76.44
P302	96.89	96.59	0.25	52.85	99.22	72.69
P303	97.68	96.73	0.25	33.83	98.99	49.09
P401	96.59	96.25	0.25	53.94	99.30	73.01
P402	97.91	96.55	0.26	27.91	98.70	38.48
P403	96.76	96.38	0.25	74.31	99.41	84.49

P501	93.31	96.23	0.25	135.58	99.54	95.47
P601	95.24	95.83	0.26	142.89	99.44	95.83
P701	95.80	97.65	0.20	76.20	99.51	88.79
P801	95.76	97.54	0.21	64.23	99.46	82.94
P901	97.77	98.34	0.18	71.25	99.47	86.70
P902	97.87	98.28	0.19	79.31	99.48	89.13
P903	97.93	98.31	0.18	78.05	99.38	89.05
P1001	62.75	97.47	0.20	50.54	99.41	77.05
P1002	97.99	98.30	0.19	70.82	99.38	86.75
P1003	97.38	98.23	0.19	74.50	99.49	87.56
P1101	98.31	98.21	0.20	68.41	99.42	84.36
P1102	98.40	98.32	0.20	58.27	99.34	79.93
P1103	98.51	98.27	0.20	64.83	99.46	83.36

Table 2.3: GATK variant calling metrics

N: number, Ti: Transition, Tv: transversion. Hom: homozygous

Sample	nVariant Loci	nNovel Variant	nSNPs	nIndels	nHom	nTi	nTv
P101	43273	592	39141	3313	16449	28370	10761
P102	42953	558	38934	3251	16726	28258	10666
P103	43465	606	39371	3298	16520	28528	10837
P201	42759	589	38831	3196	16846	28191	10628
P202	40085	491	36680	2884	16663	26676	9997
P203	42646	569	38688	3165	16792	27949	10733
P301	43017	588	39016	3232	16606	28251	10755
P302	43347	599	39318	3256	16660	28443	10865
P303	41937	531	38170	3079	16791	27645	10515
P401	43589	706	39495	3302	17037	28579	10905
P402	42590	664	38770	3165	16636	28047	10711
P403	43739	725	39572	3354	17205	28661	10897
P501	44179	1090	39876	3441	17331	28761	11101
P601	52709	876	47759	3950	18160	34522	13214
P701	43468	603	39305	3336	16919	28401	10889
P801	43100	602	38985	3319	16506	28178	10792
P901	43040	939	39082	3289	18359	28282	10780
P902	42986	907	38957	3325	18221	28210	10726
P903	43674	940	39612	3369	18418	28673	10927
P1001	44467	1215	39340	4213	16249	28491	10830
P1002	43468	696	39350	3423	16353	28488	10845
P1003	42907	648	38912	3346	16586	28096	10806

P1101	43093	941	39077	3333	18776	28175	10885
P1102	43204	858	39228	3294	18578	28372	10837
P1103	43188	903	39194	3305	18420	28237	10937

Table 2.4: SnpEff annotation metrics

Sample	frame_ shift	stop_ gained	stop_ lost	start_ lost	splice_ site	missense	inframe	synonymous	UTR
P101	227	65	23	16	65	8674	182	9718	1999
P102	220	60	22	18	58	8511	185	9686	1988
P103	213	59	24	15	71	8809	171	9710	1986
P201	204	54	27	12	62	8597	181	9684	1914
P202	191	45	24	14	57	8184	177	9325	1855
P203	215	63	24	9	58	8534	175	9551	1915
P301	209	62	22	15	52	8700	190	9685	1937
P302	209	59	20	14	52	8633	188	9763	1928
P303	195	51	25	12	52	8560	177	9525	1936
P401	210	59	26	18	61	8658	206	9862	2015
P402	220	50	26	18	62	8637	201	9862	1955
P403	218	51	25	16	60	8636	204	9856	2015
P501	231	57	24	12	62	8668	221	9960	1988
P601	236	63	26	19	67	10229	217	12000	2390
P701	206	54	23	12	65	8635	217	9713	1982
P801	210	54	25	16	64	8533	201	9770	2016
P901	210	62	25	20	70	8627	208	9809	1954
P902	240	66	24	20	64	8579	193	9688	1958
P903	232	62	24	15	72	8634	214	9903	1992
P1001	243	47	26	15	62	8644	258	9882	2070
P1002	234	56	24	19	55	8722	230	9830	1978
P1003	218	59	27	12	64	8495	209	9779	2002
P1101	231	64	22	21	63	8580	194	9700	1998
P1102	221	64	22	21	63	8558	197	9742	2020
P1103	229	50	22	17	65	8536	182	9796	1965

2.3.3 Variant analysis

Thirty-nine rare nonsynonymous variants (MAF < 1%) were observed in the 11 patients among which 12 were never seen in the same zygosity form (i.e. heterozygous or homozygous) in ExAC (Table 2.5). In addition to the above 39 variants, we identified 76 rare (MAF < 1%) nonsynonymous variants in 60 known PID genes (Table 2.6). Functional classification of these 60 genes using DAVID highlighted the complement pathway as the most enriched cluster: 13 of the 29 complement pathway genes present in the list of 251 known PID genes carried at least one rare, nonsynonymous coding variant (DAVID enrichment score: 12.57, hypergeometric probability: 0.0058).

In one patient with fatal septic shock, we identified a novel single-base insertion on, the X chromosome (C>CT at position 100617191) leading to a frame-shift in the Bruton agammaglobulinemia Tyrosine Kinase (BTK) gene, BTK plays a crucial role in B-cell development. Postmortem serological testing confirmed absence of immunoglobulins, consistent with BTK loss-of-function (IgG <1.3g/L; IgA <0.2g/L; IgM <0.2g/L).

Table 2.5: Potentially causal variants

Potentially causal variants in our 11 patients, never seen in the same zygosity form (i.e. heterozygous or homozygous) in ExAC.

Inheritance	Chromosome	Position	Gene	Effect	ExAC MAF (%)
Autosomal recessive	7	142562052	EPHB6	INSERT	0
	20	31393172	DNMT3B	DELETE	0
de novo	19	41025445	SPTBN4	MISSENSE	0
	13	112721975	SOX1	MISSTART	0
X-linked	X	27998602	DCAF8L1	MISSENSE	1.14E-03
	X	100617192	BTK	FRAMESHIFT	0
	X	9864517	SHROOM2	MISSENSE	0
	X	53575033	HUWE1	MISSENSE	0
	X	70824419	ACRC	MISSENSE	0
	X	68382133	PJA1	MISSENSE	0
	X	101097713	NXF5	DONOR	0
	X	109561080	AMMECR1	MISSENSE	0

Table 2.6: Rare nonsynonymous variants in PID genes

Chr: chromosome, n: number. *: Homozygous or hemizygous variant

Chr.	Position	Effect	Gene	MAF (%)	nHom in ExAC	nHet in ExAC
1	207646266	MISSENSE	CR2	3.30E-03	0	4
1	949431	MISSENSE	ISG15	4.17E-03	0	5
1	235896980	MISSENSE	LYST	0.02	0	27
1	154247666	MISSENSE	HAX1	0.03	0	31
1	22965341	MISSENSE	C1QA	0.04	0	13
1	11094908	MISSENSE	MASP2	0.08	1	95
1	183536358	MISSENSE	NCF2	0.23	2	273
1	949422	MISSENSE	ISG15	0.16	3	192
1	151316324	MISSENSE	RFX5	0.88	10	1045
1	207925595	MISSENSE	CD46	0.5	11	585
1	196684855	MISSENSE	CFH	0.5	17	577

1	196799796	MISSENSE*	CFHR1	0.14	41	74
1	196715063	MISSENSE	CFH	0.97	58	1056
2	231036831	MISSENSE	SP110	4.12E-03	0	5
2	47277182	MISSENSE	TTC7A	0.2	1	245
2	47273468	MISSENSE	TTC7A	0.21	2	248
3	196198925	MISSENSE	RNF168	0.2	6	232
4	151242409	MISSENSE	LRBA	0.46	2	560
4	110667485	MISSENSE	CFI	0.34	5	405
4	187004767	MISSENSE	TLR3	0.3	7	339
5	39342214	NONSENSE	C9	0.1	0	120
5	77334907	MISSENSE	AP3B1	0.08	0	97
5	1268697	MISSENSE	TERT	0.09	0	113
5	147475388	MISSENSE	SPINK5	0.65	3	779
5	35876300	MISSENSE	IL7R	0.24	5	283
5	41155088	MISSENSE	C6	0.69	8	823
5	41155088	MISSENSE	C6	0.69	8	823
5	158750329	MISSENSE	IL12B	0.65	12	763
5	40945397	MISSENSE	C7	0.95	35	1069
6	31915584	MISSENSE	CFB	0.1	0	116
6	32798457	MISSENSE	TAP2	0.49	2	569
6	137540425	MISSENSE	IFNGR1	0.14	4	156
6	32800427	MISSENSE	TAP2	0.61	8	697
8	100844596	DISRUPT	VPS13B	0.43	0	330
8	100654621	MISSENSE	VPS13B	0.02	0	20
8	100861113	MISSENSE	VPS13B	0.06	1	74
8	48733399	MISSENSE	PRKDC	0.09	2	108
8	90982691	MISSENSE	NBN	0.26	6	299
8	42177163	MISSENSE	IKBKB	0.89	59	958
9	340168	MISSENSE	DOCK8	0	0	0
9	139264888	MISSENSE	CARD9	0.35	2	406
9	139840153	MISSENSE	C8G	0.46	6	262
9	123751873	MISSENSE	C5	0.61	20	691
9	123737145	MISSENSE	C5	0.91	47	1012
10	97983635	MISSENSE	BLNK	0.55	10	645
10	73103969	MISSENSE	SLC29A3	0.54	19	617
10	6063567	MISSENSE	IL2RA	0.88	51	969
11	118898444	MISSENSE	SLC37A4	0	0	0
11	4104647	MISSENSE	STIM1	8.25E-04	0	1
11	108117787	MISSENSE	ATM	0.13	0	155

11	6637588	MISSENSE	TPP1	0.18	1	216
11	2407334	MISSENSE	CD81	0.56	1	100
11	108119823	MISSENSE	ATM	0.22	2	260
11	108129778	MISSENSE	ATM	0.21	5	247
11	108123551	MISSENSE	ATM	0.29	5	335
11	108098576	MISSENSE	ATM	0.74	5	883
11	108138003	MISSENSE	ATM	0.91	13	1084
12	122064747	MISSENSE	ORAI1	0	0	0
12	110017618	MISSENSE	MVK	0.07	0	87
12	133263886	MISSENSE	POLE	0.14	0	21
12	110034347	MISSENSE*	MVK	0.13	6	141
15	91337505	MISSENSE	BLM	0.1	1	119
15	91295110	MISSENSE*	BLM	0.86	43	956
16	27460420	MISSENSE	IL21R	0.03	0	34
16	50745960	MISSENSE	NOD2	0.03	0	41
16	81957106	MISSENSE	PLCG2	0.13	2	154
17	73826517	MISSENSE	UNC13D	2.25E-03	0	1
17	26875685	MISSENSE	UNC119	0.01	1	11
17	76120792	MISSENSE	TMC6	0.65	7	351
19	18170874	MISSENSE	IL12RB1	0.17	1	204
20	31393172	DELETE*	DNMT3B	0	0	0
20	62324328	MISSENSE	RTEL1	0.98	61	1052
22	36662063	MISSENSE	APOL1	0.01	0	18
22	31007023	MISSENSE	TCN2	0.27	6	314
X	100617192	FRAMESHIFT	BTK*	0	0	0
X	77150892	MISSENSE	MAGT1	0.29	106	120

A novel homozygous deletion spanning 6 base pairs on chromosome 20 (ACTCGAG>A at position 31393171) was observed in a two-year-old boy with consanguineous parents leading to an in-frame deletion in a conserved region of the catalytic domain of the DNMT3B gene. Previously described missense mutations in the same protein domain are known to cause immunodeficiency-centromeric instability-facial anomalies (ICF) syndrome (*11*), a rare disease characterized by variable immunodeficiency and recurrent infections with mild facial abnormalities. The fatal septicaemia was preceeded by mild varicella and the patient had a history of recurrent mild respiratory and ENT infections that had been attributed to poor health conditions. The patient had normal T and B cell subsets but reduced immunoglobulin levels, consistent with ICF syndrome: IgG 0.16g/L (4.22-11.9); IgA <0.06g/L (0.2-1.58); IgM 0.05g/L (0.48-1.9).

We also observed a known, very rare (MAF=0.00015 in ExAC) heterozygous stop-gain variant in the complement C9 gene, on chromosome 5 (G>T at position 39342214) in a male patient with non-fatal *Pseudomonas* septicaemia and recurrent parainfectious neutropenia. However, complement reconstitution assay with classical or terminal pathway protein (C1 to C9) showed normal lytic activity (Figure 2.2), suggesting that this variant is unlikely to be causally related to increased susceptibility to *P. aeruginosa*.

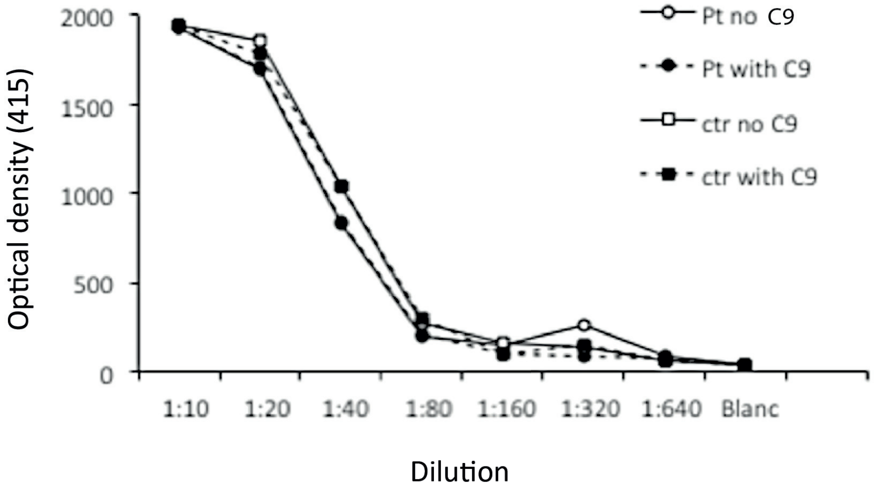

Figure 2.2: Serum lytic activity.

The stop gain mutation in C9 does not affect the lytic activity of patient's serum compare to control.

2.4 Conclusion and discussion

While mortality due to bacterial sepsis has decreased over the past decade, one out of three pediatric sepsis deaths occur in children without major comorbidities (*12*). Invasive infections in children in whom immunodeficiency had not been previously suspected may represent the first manifestation of a PID. However, very little is known about the prevalence of PIDs in children presenting with bacterial sepsis. We used exome sequencing to explore the genetic cause of susceptibility to *P. aeruginosa* in a selected group of previously apparently healthy children with community-acquired *P. aeruginosa* septicemia. We identified PIDs listed in the 2014 International Classification of primary immunodeficiencies in two cases: X-linked agammaglobulinemia due to a novel *BTK* mutation and ICF immunodeficiency syndrome due to a novel *DNMT3B* mutation. In addition, we observed 76 rare missense variants in known PID genes, in particular within the complement pathway. Functional testing of a surviving patient with a heterozygous protein-truncating variant in *C9* revealed normal lytic activity, demonstrating the need for strict validation of potential causal mutations, even in genes with a plausible biological link to the study phenotype.

Previous case reports describing PIDs such as Wiskott-Aldrich syndrome, X-linked agammaglobulinemia, cyclic neutropenia and IRAK-4/MyD-88 deficiency (*2*, *3*, *13*) in *P.*

aeruginosa sepsis were based on conventional candidate-based immunologic testing. In view of the genetic heterogeneity in human susceptibility to sepsis, classic PID screening may fail to identify new mutations in known PID genes and makes no attempt at expanding our understanding of the genetic causes of the disease, with potentially devastating consequences for survivors, undiagnosed siblings and their families.

In our cohort, defects in known PID genes were found in 25% of fatal cases. While non-genetic factors including pathogen variability and non-genetic causes of neutropenia may be responsible for the remaining cases, we cannot rule out that mutations in genes not previously associated with PID are at least partially involved. Sequencing of additional family members, sequencing of more cases with the same phenotype and functional characterization of new candidate genes and variants will be required to investigate these genetic factors.

In conclusion, this study provides proof-of-concept that exome sequencing allows the identification of rare, deleterious human genetic variants responsible for fulminant sepsis in children in whom immunodeficiency had not been previously suspected. We were able to diagnose PID in several cases, with a significant impact on affected families. Given the decreasing cost of exome sequencing, we propose to consider it as a diagnostic procedure for apparently healthy children presenting with unusually severe or fatal sepsis.

2.5 References

1. Y.-C. C. Huang, T.-Y. Y. Lin, C.-H. H. Wang, Community-acquired Pseudomonas aeruginosa sepsis in previously healthy infants and children: analysis of forty-three episodes., *Pediatr. Infect. Dis. J.* **21**, 1049–52 (2002).
2. S. N. Wong, A. Y. Tam, R. W. Yung, E. Y. Kwan, N. N. Tsoi, Pseudomonas septicaemia in apparently healthy children., *Acta Paediatr Scand* **80**, 515–20 (1991).
3. T. Stergiopoulou *et al.*, Deficiency of interleukin-1 receptor-associated kinase 4 presenting as fatal Pseudomonas aeruginosa bacteremia in two siblings., *Pediatr. Infect. Dis. J.* **34**, 299–300 (2015).
4. C. Picard *et al.*, Clinical features and outcome of patients with IRAK-4 and MyD88 deficiency., *Medicine (Baltimore)* **89**, 403–25 (2010).
5. J.-L. L. Casanova, L. Abel, Primary immunodeficiencies: a field in its infancy., *Science* **317**, 617–9 (2007).
6. B. Goldstein, B. Giroir, A. Randolph, International pediatric sepsis consensus conference: definitions for sepsis and organ dysfunction in pediatrics., *Pediatr Crit Care Med* **6**, 2–8 (2005).
7. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform., *Bioinformatics* **25**, 1754–60 (2009).
8. P. Cingolani *et al.*, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3., *Fly* **6**, 80–92 (2012).
9. W. Al-Herz *et al.*, Primary immunodeficiency diseases: an update on the classification from the international union of immunological societies expert committee for primary immunodeficiency., *Front Immunol* **5**, 162 (2014).
10. D. W. a W. Huang, B. T. Sherman, R. A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources., *Nat Protoc* **4**, 44–57 (2009).
11. C. M. Weemaes *et al.*, Heterogeneous clinical presentation in ICF syndrome: correlation with underlying gene defects., *Eur. J. Hum. Genet.* **21**, 1219–25 (2013).
12. L. J. Schlapbach *et al.*, Mortality related to invasive infections, sepsis, and septic shock in critically ill children in Australia and New Zealand, 2002-13: a multicentre retrospective cohort study., *Lancet Infect Dis* **15**, 46–54 (2015).
13. M. Baro, M. A. Marín, J. Ruiz-Contreras, S. F. de Miguel, I. Sánchez-Díaz, Pseudomonas aeruginosa sepsis and ecthyma gangrenosum as initial manifestations of primary immunodeficiency., *Eur. J. Pediatr.* **163**, 173–4 (2004).

Chapter 3 Loss-of-function mutations in IFIH1 predispose to severe viral respiratory infections in children

3.1 Inroduction

Viral respiratory tract infections are the most common pediatric infections worldwide. During early childhood, virtually every individual goes through several episodes of viral respiratory infections, which are usually mild and self-limiting. Yet, serious disease leads to the hospitalization of about 1% of individuals in each birth cohort. In-hospital mortality rates are limited to 0.5% with intensive care support, still these infections account for 21% of childhood mortality worldwide (*1*, *2*). The main viral pathogens causing lower respiratory tract infections are human respiratory syncytial virus (RSV), enteroviruses (including rhinoviruses), adenoviruses, metapneumovirus, influenza and parainfluenza viruses, with RSV being responsible for the majority of the hospitalized cases (*3*, *4*). A number of risk factors including socio-economic and environmental influences (e.g. malnutrition and smoking parents), male gender, preterm birth, chronic diseases and immunosuppression are associated with more severe disease (*5*, *6*). Yet, approximately 1 out of 1000 children without any known risk factor or co-morbidity will require intensive care support due to life-threatening manifestations of common viral respiratory infections. This “silent epidemic” of potentially lethal infections before reproductive age remains largely unexplained. In the absence of established differences in pathogen virulence, we hypothesized that human genetic variation could be involved. We therefore combined exome and transcriptome sequencing and *in vitro* functional testing to identify and characterize potentially causal genetic variants in carefully selected clinical cases.

3.2 Material and methods

3.2.1 Subject recruitment and specimen collection

Previously healthy children suffering from severe lower respiratory tract infections were recruited between 1.12.2010 and 30.09.2013 in five specialized Pediatric Intensive Care Units (PICU) in Australia and Switzerland (Australia: Mater Children’s Hospital, and Royal Children’s Hospital, Brisbane; Switzerland: Children’s Hospital Lucerne, Lucerne; Centre Hospitalier Universitaire Vaudois, Lausanne; Département de l’Enfant et de l’Adolescent, Hospital Universitaire de Genève, Geneva). The study was approved by the institutional Human Research Ethics Committees. Informed consent was obtained from parents or legal guardians.

Children aged below 4 years that were admitted to PICU due to a severe respiratory infection of proven or presumed viral origin, and who required respiratory support were eligible. Exclusion criteria were the presence of any significant underlying disease or comorbidity, including prematurity, congenital cardiac disease, chronic lung disease including cystic fibrosis, sickle cell disease, hepatic, renal, or neurologic chronic conditions, solid and haematological malignancies and known PID. Respiratory support was defined as non-invasive modes of ventilation (NIV) including high-flow nasal cannulae (HFNC) and continuous or bilevel positive airway pressure (CPAP and BiPAP), vs. invasive (IV) ventilation including conventional and high-frequency oscillation ventilation (HFOV). The following demographic and clinical information was collected for each recruited subject: age, gender, weight, ethnicity, type of ventilation needed, length of ventilation in days, clinical outcome, microbiological diagnostic procedures and results including rapid test for RSV and influenza, viral PCR, and viral cultures. The Pediatric Index of Mortality 2 (PIM2) was used to assess patient illness severity at PICU admission.

Parents were approached by the study investigator immediately following PICU admission, and, when consent was available, blood and respiratory samples were obtained. For each study subject a nasopharyngeal aspirate, 1ml EDTA blood in vacutainer tubes and 2.5ml blood in PAXgene blood RNA tubes (catalog number 762165; PreAnalytiX, Hombrechtikon, Switzerland) were collected. Samples were immediately frozen at -70 degrees Celsius until shipment, and then analyzed in batch.

3.2.2 Screening of respiratory viruses

Samples were extracted individually (100ul sample from nasopharyngeal aspirate + 300ul PBS) using the NucliSens Easymag© (bioMérieux, Geneva, Switzerland) according to the manufacturer's instructions. Respiratory viruses screening was performed using FTD Respiratory pathogens 21 assay (Fast-track Diagnostics) on a Viia7 instrument (Applied biosystems, Switzerland).

3.2.3 Exome sequencing and alignment

Genomic DNA was extracted from whole blood (catalog number 51104; QIAGEN, Hombrechtikon, Switzerland). Exome sequencing libraries were prepared with 2 µg to 3 µg of genomic DNA using Agilent SureSelect reagents (catalog numbers 5190-4627, 5190-4631, 5190-6208 and G9611A; Agilent Technologies, Santa Clara, USA) according to the protocols supplied by the manufacturer. Cluster generation was performed using the Illumina TruSeq PE Cluster Kit v3 reagents. The resulting libraries were sequenced as 100-nucleotide, paired-end reads on Illumina HiSeq 2000 or HiSeq 2500 using TruSeq SBS Kit v3 reagents. Sequencing was done at the Lausanne Genomic Technology Facility. Raw sequencing reads were processed using the Illumina Pipeline Software version 1.82. Purified filtered reads were aligned to human reference genome hg19 using Burrow-Wheeler Aligner version 0.6.2 (BWA) (*7*) (supplementary table 2). PCR duplicates were removed using Picard (http://picard.sourceforge.net/).

3.2.4 Variant calling

We used Genome Analysis Toolkit (GATK) (*8*, *9*) version 3.1-1, Platypus (*10*) version 0.7.9.1 and SAMtools (*11*) version 0.1.19 to call single nucleotide variants (SNVs) and small insertion and deletions (Indels) from duplicate-marked bam files. With GATK we used HaplotypeCaller for multi-sample variant calling. We followed GATK best practice to call the variants (https://www.broadinstitute.org/gatk/guide/best-practices?bpm=DNAseq); variants that didn't pass GATK filtering were filtered out. With Platypus we used callVariants for multi-sample variant calling and variants that didn't pass Platypus filtering criteria were filtered out. With SAMtools we used mpileup default options for multi-sample variant calling followed by bcftools view to generate the vcf file, variants with quality score below 30 were filtered out. After variant calling, variants that were present in the intersection of the three variant callers with missingness below 5% were used for further annotation and down stream analysis.

3.2.5 Variant annotation

We used SnpEff (*12*) version 4.1B to predict the functional impact of variants, and notably to identify putative loss of function variants (LoFs). We defined these categories of variantions as LoF: stop-gain SNVs, splice site disrupting SNVs and frame shift indels in the first 95% of coding region or larger deletions removing either the first exon or more than 50% of the protein-coding sequence of the affected transcript. We further enriched for LoFs that are likely to abolish protein structure by including only LoFs that affected all the coding transcripts of a gene.

3.2.6 RNA sequencing and alignment

Total RNA was extracted from 2.5 ul of whole blood using PAXgene blood RNA kit (catalog number 762174; PreAnalytiX, Hombrechtikon, Switzerland). Libraries were prepared using the Illumina TruSeq Stranded mRNA (Ribo-Zero Globin) reagents (Catalog number RS-122-2501; Illumina; San Diego, USA) according to the protocol supplied by the manufacturer and using 400ng of total RNA. Cluster generation was performed with the Illumina TruSeq PE Cluster Kit v3 reagents. The resulting stranded libraries were sequenced as 100-nucleotide, paired-end reads on the Illumina HiSeq 2000 using TruSeq SBS Kit v3 reagents. The raw sequencing reads were processed using the Illumina Pipeline Software version 1.82. Purified filtered reads were aligned to the human reference genome hg19, using STAR (*13*) version 2.3.0e and the Gencode annotation (*14*) version 19.

3.2.7 Plasmids

β-IFN-fl-lucter contains the firefly luciferase gene driven by the human IFNβ promoter as described previously (*15*). pTK-rl-lucter contains the *Renilla* luciferase gene (PROMEGA) driven by the herpes simplex virus TK promoter. pEBS-tom encodes a red fluorescent protein.

pcDNA3.1(+) containing wild-type (wt) *IFIH1* was constructed by PCR amplification on pEF-BOS-IFIH1 with sense primer that introduced a BamHI site and flag sequence at the

N-terminus and with the antisense primer that introduced a XhoI site at the C-terminus of IFIH1. The PCR products were inserted into pcDNA3.1(+) after digestion with BamHI and XhoI.

pcDNA3.1(+) containing mutant *IFIH1* lacking exon 14 *(IFIH1-Δ14),* was constructed using a fusion PCR strategy. First, pEF-BOS-IFIH1 was amplified by PCR with a sense primer that introduced a BamHI site and a HA sequence at the N-terminus of *IFIH1* and with the antisense primer that inserted the beginning of exon 15 sequence at the end of the exon 13. Second, pEF-BOS-IFIH1 was amplified by PCR with a sense primer that introduced the end of exon 13 sequence at the beginning of exon 15 and with the antisense primer that introduced a XhoI site at the C-terminus of *IFIH1*. The resulting PCR products were mixed and amplified by PCR with a sense primer that introduced a BamHI site and a HA sequence at the N-terminus and with the antisense primer that introduced a XhoI site at the C-terminus of IFIH1. The PCR products were digested with BamHI and XhoI and then inserted into pcDNA3.1(+).

pcDNA3.1(+) containing deletion mutant *IFIH1* lacking exon 8 *(IFIH1-Δ8),* and pcDNA3.1(+) containing deletion mutant *IFIH1* with a stop-gain mutation in exon 10 *(IFIH1-ΔCTD),* were ordered from life technologies plasmid service.

The inserts of the resulting pcDNA3.1(+) containing *IFIH1-wt* or mutant plasmids were confirmed by sequencing.

3.2.8 Transfection & measure of interferon-β promoter activity

293T cells and Huh7.5 cells were grown in Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum and 1% Pen/Strep. 100'000 cells were plated into 6-well plates and transfected 24 hrs later with 1.5 µg of pβ-IFN-fl-lucter, 0.5 µg of pTK-rl-lucter and 0.5 µg of pEBS-tom (used as a transfection control), using Gene Juice transfection reagent (NOVAGEN). Additionally, cells were transfected with 1 µg of IFIH1 encoding plasmids. 24 hrs later, Huh7.5 cells (but not 293T cells) were transfected with elicitor RNA using TransMessenger transfection reagent (QIAGEN) according to the manufacturer's instructions. 20 hrs later, cells were harvested and cell lysates were used to measure firefly and *Renilla* luciferase activity (dual-luciferase reporter assay system, Glomax 20/20 luminometer, PROMEGA).

Protein expression in the cell lysates was then checked by Western blotting, using the following primary antibodies: anti-RIG-I (1:1000) (Alexis ALX-210-932), anti-IFIH1 (1:1000) (Alexis ALX-210-935), anti-HA (1:2000) (Enzo Life Sciences ENZ-ABS-118-0500), anti-flag (1:1000) (Sigma F1804-1MG). Immunoblot analyses were developed with the following secondary antibodies: goat anti-mouse and anti-rabbit IgG horseradish peroxidase conjugated whole antibody (1:3000) (Bio-Rad). ImageJ version 1.44p (http://imagej.nih.gov/ij/) was used for western blot quantifications.

3.2.9 Viruses

Respiratory Syncytial Virus expressing Green Fluorescent Protein (RSV-GFP, obtained from Mark peeples (*49*)) or mCherry (rRSV-mCherry obtained from Jean-François Eléouët (*50*)) stock were amplified in A549 (human alveolar adenocarcinoma cell line) cells. RSV-GFP stock titers were determined using serial dilutions to infect A549 cells. GFP or mCherry expression was measured by flow cytometry on 20’000 cells (*i.e.* 10% of the harvested cells) using BD Accuri C6 Cytometer. Data were analyzed using CFlow Plus software (Accuri, version 1.0.264.15). Fluorescence pictures were acquired using Evos FL epifluorescence microscope. Recombinant vesicular stomatitis virus expressing Green Fluorescent Protein (rVSV-GFP obtained from Jacques Perrault (*51*)) stock were amplified in 293T cells and stock titers were determined using serial dilutions to infect 293T cells. GFP expression was measured by flow cytometry on 20’000 cells (*i.e.* 10% of the harvested cells) using BD Accuri C6 Cytometer. Data were analyzed using CFlow Plus software (Accuri, version 1.0.264.15). Fluorescence pictures were acquired using Evos FL epifluorescence microscope.

3.2.10 Measure of protein stability by pulse chase

Transfected 293T cells were incubated for 30 minutes at 37°C in methionine-free, cysteine-free and FCS-free DMEM. 100 µCi/ml of 35S-methionine + 35S-cysteine labelling mix (HARTMAN Analytic) was added and cells were incubated at 37°C for 30 minutes. The chase (0, 2, 4, and 8 hours) was performed at 37°C in DMEM supplemented with unlabelled methionine and cysteine (10mM). Cell lysates were loaded on 7.5% acrylamide gels, transferred to a PVDF membrane and exposed to autoradiography. Results were revealed in a phosphorimager (Typhoon, GE Healthcare Life Sciences) and quantified with ImageQuantTL software (GE Healthcare Life Sciences).

3.2.11 Recombinant IFIH1 expression

IFIH1 inserts cloned into pcDNA3.1(+) were inserted into pET28-His10Smt3 backbone. pET28-His10Smt3-IFIH1 wt or ΔCTD plasmids were transformed into *E. coli* BL21. Cultures (500 ml) derived from single transformants were grown at 37°C in LB medium containing 50 *µ*g/ml kanamycin to an A_{600} of 0.6. The cultures were adjusted to 0.2 mM IPTG and 2% ethanol and further incubated for 20 hours at 17°C. Cells were harvested by centrifugation and recombinant RIG-I protein was purified from bacteria as previously described (*8*). Protein concentration was determined using the Bio-Rad dye binding method with BSA as the standard.

3.2.12 Measure of IFIH1 ATPase activity

Increasing amounts of polyI:C were incubated with 200 nM of purified recombinant IFIH1, [(Isqb) γ-^{32}P] ATP (Hartmann Analytic) in a final volume of 15 µl (50 mM Tris acetate pH 6, 5 mM DTT, 1.5 mM $MgCl_2$) for 15 minutes at 37°C. Reactions were then stopped with 1 mM formic acid and 2.5 µl of each reaction were spotted onto TLC PEI Cellulose F plates (MERCK 1.05579.0001) and applied to a migration buffer (LiCl 0.5 M and formic acid 1 N) to separate released $^{32}PO_4$ and non hydrolyzed ATP. $^{32}PO_4$ release was measured in a phosphorimager (Typhoon, GE Healthcare Life Sciences) and quantified with

ImageQuantTL software (GE Healthcare Life Sciences). ATPase data were processed as follow: for each sample, the ratio free $^{32}PO_4$/non hydrolyzed ATP was calculated (and reported to the final ATP concentration in the reaction); fold increases were obtained by normalizing the calculated ATPase activity to the ATPase activity of the protein alone (control without RNA).

3.2.13 Cell culture and transduction with lentivirus vectors

The 293T cells line were cultured in Dulbecco's modified Eagle's medium supplemented with 10% fetal calf serum. All recombinant lentiviruses were produced by transient transfection of 293T cells according to the following protocol. 5 x 10^6 293 cells were plated in 10 cm PD and cotransfected with 15 µg of a plasmid vector (pLV.CMV.IFIH1.IRES-GFP and pLV-U6-empty-PGK-GFP), 10 µg of psPAX2 and 5 µg of pMD2G-VSVG by calcium phosphate precipitation. After 8 hours, medium was changed and recombinant lentiviruses vectors were harvested 24 hours later. Huh 7.5 cell lines, platted in 6-well plates, were transduced at an MOI of 2 with the recombinant lentiviruses. 2 days after transduction, GFP-expressing cells were sorted by FACS.

3.2.14 Quantitative RT-PCR

Total RNA isolation was performed by using the NucleoSpin RNA kit (Macherey-Nagel, Switzerland) according to the manufacturer’s instruction. 200 µg of total RNA from infected cells pellet was used for cDNA synthesis using the Omniscript RT Kit (Qiagen, Basel, Switzerland).

Quantitative RT-PCR was performed by using 2 µL of cDNA and 18 µL of Taqman Fast universal PCR master mix (Thermo Fisher Scientific, Waltham, MA, USA) containing specific primers (20 µM) and probes (5 µM) for human rhinovirus (HRV). 7500 fast Real-Time PCR System (Thermo Fisher Scientific) was used to perform PCR reactions. The ΔΔCT method was used to quantify the mRNA expression levels of endogenous genes. The mRNA expression levels of endogenous genes were normalised to the housekeeping gene 18S rRNA. Viral RNA copy numbers were obtained through the generation of a standard curve obtained with serial dilutions of a plasmid containing HRV cDNA.

3.2.15 Statistical analysis

Unpaired T-tests were performed using R (version 1.65).

3.3 Results

3.3.1 Clinical presentation and outcomes

We prospectively enrolled 120 previously healthy children admitted to an intensive care unit (ICU) for ventilation support due to a common viral respiratory infection. Children with known risk factors (prematurity, chronic disease, immunosuppression) were excluded. After informed consent was obtained from parents or legal guardians, we collected detailed clinical information and laboratory results, as well as biological samples for human DNA and RNA analysis (blood) and for viral diagnostics (endotracheal tube aspirate or nasopharyngeal aspirate). The most common clinical presentation was bronchiolitis (88%) and the median age was 78 days (range 7 – 828). RSV was the most common pathogen, identified in 56% of cases, followed by enteroviruses (rhinovirus / enterovirus) (Table 3.1).

Table 3.1: Summary of patients' demographic information

Baseline characteristics of patients with severe viral respiratory infections admitted to pediatric intensive care unit (PICU). RSV: respiratory syncytial virus, CoVHKu1: Cronavirus Hku1, hPIV1: human Parainfluenza virus type 1, CoVnL63: Cronavirus nL63. CPAP: machine continuous positive airway pressure, BiPAP: Bilevel Positive Airway Pressure, HFOV: High Frequency Oscillatory Ventilation.

Age (days)		78 (37 – 269)
Weight (kg)		5.9 (4.4 – 9.8)
Country of recruitment	Australia	100 (83%)
	Switzerland	20 (17%)
Ethnicity	Caucasian	90 (78%)
	African	4 (4%)
	Asian	4 (4%)
	Australian aboriginal	6 (5%)
	Pacific Islander	11 (10%)
Sex	Male	70 (58%)
	Female	50 (42%)
Clinical phenotype	Bronchiolitis	105 (88%)
	Pneumonia	8 (7%)
	Laryngotracheobronchitis	5 (4%)
	Reactive airway disease	2 (2%)
Virus identified in respiratory sample	RSV	67 (56%)
	Entero/Rhinovirus	31 (26%)
	Adenovirus	17 (14%)
	Human Bocavirus	9 (8%)
	Influenza	2 (2%)
	Parainfluenza	6 (5%)
	Human Metapneumovirus	3 (2%)
	CoV-HKU1	3 (2%)
	CoV-NL63	1 (1%)
Respiratory Support	High-Flow Nasal Cannulae	99 (83%)
	CPAP or BiPAP	30 (25%)
	Invasive ventilation	31 (26%)
	HFOV	6 (5%)
Length of PICU stay (days)		2.7 (1.6 – 5.0)
Expected mortality	Pediatric index of mortality 2 (%)	0.02 (0.01 – 0.7)
Observed mortality	Pediatric index of mortality 2 (%)	No fatal case

3.3.2 Exome sequencing and alignment of short reads

For each sample, on average 93% of pass filter reads were unique (not marked as duplicate). 96% of pass filter unique reads were aligned to the human reference genome. The mean on-bait coverage was 70x with 96% of target bases achieving at least 10x coverage and 78% achieving at least 30x coverage (Table 3.2).

Table 3.2: Summary of BWA alignment metrics

SD: standard deviation

	PF unique reads (%)	PF unique reads aligned (%)	Off-bait (%)	Mean bait coverage (%)	At least 2X coverage (%)	At least 30X coverage (%)
Average	0.93	0.96	0.22	69.99	0.99	0.78
SD	0.07	0.04	0.03	15.29	0.01	0.08
Median	0.96	0.97	0.23	68.04	0.99	0.79

3.3.3 Variant calling and annotation

To reduce the number of false positive variants we called variants using three different callers: GATK, Platypus and SAMtools. Each callset was annotated and filtered separately and only the variants that were called by the three callers and had a missingness rate below 5% were kept for downstream analysis. In this final callset we identified 232,591 SNVs and 103,63 indels including: 951 frame-shift indels, 793 stop-gained and 297 splice-site variant (Table 3.3).

Table 3.3: Variant calling metrics

Variant caller	GATK	Platypus	SAMtools	Intersection
Total variants (%novel)	287681 (16.74%)	271978 (17.72%)	317820 (18.90%)	242954 (14.14%)
SNV (%novel)	263507 (14.04%)	258548 (16.34%)	296161 (16.63%)	232591 (13.25%)
indel (%novel)	24174 (46.23%)	13430 (44.34%)	21659 (49.89%)	10363 (34.12%)
FRAME-SHIFT	1644	1333	1415	951
STOP-GAINED	880	942	1082	793
STOP-LOST	89	85	99	76
START-LOST	126	128	144	112
SPLICE_SITE	377	353	446	297
MISSENSE	64057	66559	74945	59144
INFRAME	1854	1102	1811	867
SYNONYMOUS	55698	56964	62381	52261

UTR	23478	16320	19780	14434
LoF (%novel)	2635 (49.44%)	2419 (49.89%)	2710 (46.60%)	1878 (41.64%)
SNP-LoF (%novel)	1116 (30.73%)	1219 (34.12%)	1371 (32.31%)	1023 (30.00%)
Indel-LoF (%novel)	1519 (63.19%)	1200 (65.91%)	1339 (61.23%)	855 (55.55%)

3.3.4 Loss of function variants

Because they are under strong purifying selection, most LoFs have low frequencies in the population. 903 LoFs were found in the intersection of the three callsets. We focused on 568 of them (456 SNV and 103 indels), which were found at low frequency (MAF<5%) in publicly available databases and in an in-house set of 485 exomes. Only 10 of those were observed in homozygous form in our study population, impacting 10 different genes (Table 3.4).

Table 3.4: Completely knocked-out genes

MAF threshold: <= 0.01, Chr: chromosome, NA: not available, het: heterozygous.

Chr	Position	ID	Gene	Effect	PRI AF (%)	ExAC MAF (%)	ExAC het
1	26301080	rs375730731	PAFAH2	STOP-GAINED	2.083	0.013	16
2	163124596	rs35732034	IFIH1	SPLICE-SITE	2.083	0.7	813
2	186668962	NA	FSIP2	FRAME-SHIFT	0.833	0.005	1
11	18497187	rs145833201	LDHAL6A	SPLICE-SITE	0.833	0.094	112
17	79974815	rs143008575	ASPSCR1	SPLICE-SITE	0.833	0.087	102
20	31659102	rs146981368	BPIFB3	SPLICE-SITE	5.417	0.288	210
X	7889837	NA	PNPLA4	STOP-GAINED	0.833	0.002	1
X	127185638	NA	ACTRT1	FRAME-SHIFT	2.083	0.199	170

3.3.5 IFIH1 role in innate immunity

IFIH1 encodes a RIG-I-like cytoplasmic sensor of long double-stranded RNA (dsRNA) and plays a chief role in innate immune recognition of RNA viruses (*7*, *8*). It has been shown previously that IFIH1, in combination with RIG-I, recognizes and limits the replication of many RNA viruses including: positive single-stranded RNA (ssRNA) viruses especially picornaviruses (*12*, *14–21*), negative single-stranded ssRNA viruses including paramyxoviruses (*13*, *22–25*) and double-stranded RNA (dsRNA) viruses (*26*).

3.3.6 Description of *IFIH1* LoF carriers

8 patients in our cohort carried IFIH1 LoF mutations (Table 3.6). A 16-month old previously healthy girl was identified to be homozygous for rs35732034. She presented with RSV-positive respiratory tract infection requiring invasive ventilation. The disease course was complicated by superinfection with S. aureus. She showed full recovery and did not require further hospitalization outside this event up to age of three years. Her phenotype and history was otherwise unremarkable. She showed normal neurodevelopment and has not demonstrated any complication with vaccines. A basic immunologic workup including full blood count, immunoglobulin levels and IgG subclasses, and lymphocyte subclass counts resulted within normal limits. Three infants requiring non-invasive respiratory support for bronchiolitis at early infancy were found to be heterozygous for *IFIH1* rs35732034. One of these manifested recurrent severe viral lower respiratory tract infections leading to repeated PICU admissions during early childhood. Three infants (age 1 to 16 months) were found to be heterozygous for *IFIH1* rs35337543 and required non-invasiave ventilatory support for bronchiolitis due to RSV, including one with RSV and HBoV coinfection. One infant requiring invasive ventilation due to enterovirus positive bronchiolitis was heterozygous for *IFIH1* rs35744605.

Table 3.5: LoF mutations found in *IFIH1*

Chr: chromosome, hom: homozygous

Chr	Position	ID	Effect	ExAC MAF (%)	PRI AF (%)	AC (hom)
2	163124596	rs35732034	SPLICE-SITE	0.7	2.08	5 (1)
2	163134090	rs35744605	STOP-GAIN	0.33	0.42	1 (0)
2	163136505	rs35337543	SPLICE-SITE	0.65	1.25	3 (0)

3.3.7 Description of *IFIH1* LoF mutations

Four study participants carried a rare *IFIH1* splicing variant, rs35732034: one homozygous and three in heterozygous. We used RNA sequencing to characterize the transcriptomic impact of this variant. The minor allele T at rs35732034 causes skipping of exon 14 (IFIH1-Δ14), which results in a frame shift and an early stop codon in exon 15. The resulting protein lacks the final 67 amino acids of wild-type IFIH1, including the C-terminal regulatory domain (CTD), which is essential for binding to viral dsRNA. We identified two additional rare LoF variants in *IFIH1*, only present in heterozygous form in a total of four study participants: a splicing variant rs35337543 (N=3) and a stop-gained variant rs3574460 (N=1). RNA sequencing showed that the minor allele G at rs35337543 causes skipping of exon 8 (IFIH1-Δ8), which removes 37 amino acids at the end of the conserved helicase 1 domain and in the linker part between helicase 1 and helicase 2, but does not cause a frame shift. rs35744605 is a stop-gained SNV in exon 10 that leads to the loss of 398 amino acids from the C-terminal end of IFIH1 (IFIH1-ΔCTD) (Figure 3.1 A-C).

3.3.8 RNA sequencing and read mapping of IFIH1 LoFs

RNA-seq was performed on 24 biologically independent patient samples, including the sample homozygous for rs35732034, two that were heterozygous for this variant and two samples heterozygous for rs35337543. The libraries were sequenced to a depth of 98.4 million paired-end 100 base-pair reads on average (standard deviation (SD): 16.3, range: 59.5 – 121.4). An average of 84.1 million reads (SD: 13.9, range: 51.1 – 107.9) could be uniquely mapped to the reference annotation (Table 3.5).

Table 3.6: RNA-sequencing mapping metrics

Sample	Uniquely mapped (%)	Mapped to multiple positions (%)	Unmapped (%)
PRI-019	90.10	5.48	7.32
PRI-020	90.71	9.46	4.51
PRI-022	107.90	8.24	5.29
PRI-024	93.79	9.26	4.64
PRI-025	64.80	3.69	2.71
PRI-030	90.47	7.99	4.07
PRI-031	78.59	8.94	6.06
PRI-032	77.96	9.11	5.76
PRI-034	102.91	7.98	5.22
PRI-038	83.26	5.97	7.43
PRI-042	96.83	7.54	6.78
PRI-046	73.44	5.78	5.33
PRI-047	81.99	27.25	8.44
PRI-049	95.66	6.59	9.49
PRI-050	79.92	7.00	6.92
PRI-051	84.27	8.51	7.76
PRI-060	70.86	7.15	4.49
PRI-061	51.11	4.91	3.48

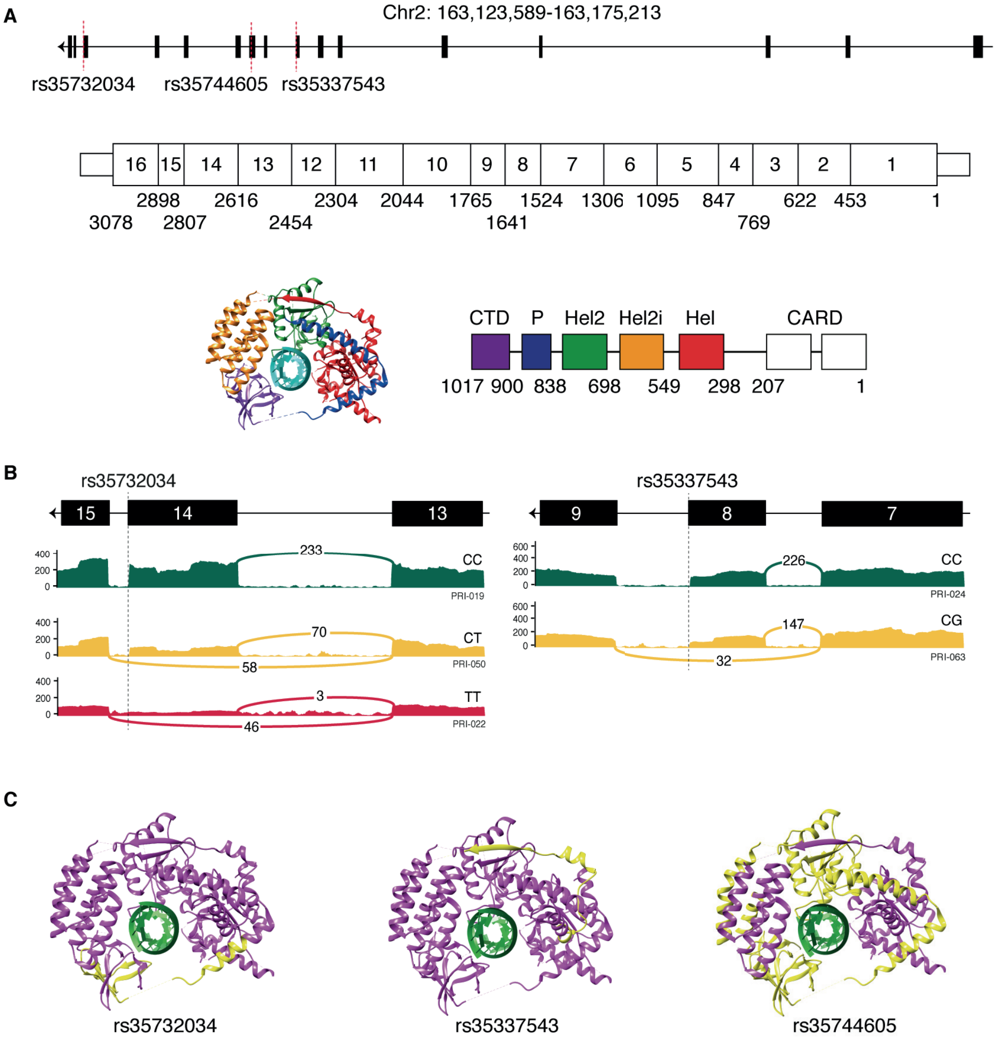

Figure 3.1: Description of *IFIH1* LoF mutations

(A) *IFIH1* gene, mRNA and protein (linear and 3D structure). Loss of function variants identified in IFIH1 are marked on IFIH1 gene. Exon boundaries are marked with nucleotide coordinates. Protein domain boundaries are marked with amino acid coordinates. CARD, caspase activation recruitment domain; Hel, helicase domain; P, pincer; CTD, C-terminal regulatory domain. (B) Alternative splicing of IFIH1 associated with rs35732034 and rs35337543 genotypes, seen in RNA sequencing data. The T allele at rs35732034 leads to exon 14 skipping. The G allele at rs35337543 leads to exon 8 skipping. The Sashimi plots illustrate the genotype-dependent abundance of splice junctions. The number of observed reads spanning the respective splice junctions is indicated on the Bezier curves, which connect exons. (C) Schematic three-dimensional representation of the IFIH1 structure (PDB: 4GL2, image produced using UCSC Chimera). The parts of the protein that are predicted to be missing due to rs35732034, rs35744605 and rs35337543 mutations are indicated in yellow.

3.3.9 Functional characterization of IFIH1 LoF mutations

To functionally characterize the identified variants, we first measured the ability of wild type (IFIH1-wt) and mutant IFIH1 isoforms to induce interferon β (IFNβ) *in vitro*. We transfected plasmids carrying IFIH1-wt, IFIH1-Δ8, IFIH1-Δ14 and IFIH1-ΔCTD into 293T cells, Overexpression of IFIH1-wt, but not of any of the mutant IFIH1 isoforms, led to IFNβ induction. Co-transfection of IFIH1-wt with each of the mutant IFIH1 isoforms showed a slight interference with IFIH1-wt induced IFNβ production (Figure 2A). We then checked the stability of the various IFIH1 protein isoforms by performing pulse-chase experiments in transfected 293T cells. The three mutant isoforms were less stable than IFIH1-wt (Figure 2B) and had a negative impact on the stability of the wild type isoform when co-transfected (Figure 2C). Finally, we tested the ATPase activity of recombinant IFIH1-wt and mutant IFIH1 isoforms. IFIH1-wt was able to hydrolyze ATP and showed a typical dsRNA-dependent increase in enzymatic activity, while the mutant isoforms had no detectable activity, even upon stimulation with polyinosinic:polycytidylic acid (polyI:C), a synthetic analogue of double-stranded RNA (Figure 2D). Furthermore, all mutant isoforms decreased IFIH1-wt ATPase activity in a dose-dependent manner, an interference that was specific to mutant IFIH1 isoforms, as demonstrated by the absence of any effect of bovine serum albumin on IFIH1-wt ATPase activity (Figure 2E-F). Jointly, these experiments demonstrate that the three putative LoF variants identified in our study population lead to severe disruption of IFIH1 signalling function, protein stability and enzymatic activity *in vitro*. In addition, the observations that the mutant IFIH1 isoforms interfere with the wild type protein in terms of IFNβ induction, protein stability and enzymatic activity suggest a potential dominant negative role for heterozygous LoF variants in *IFIH1*.

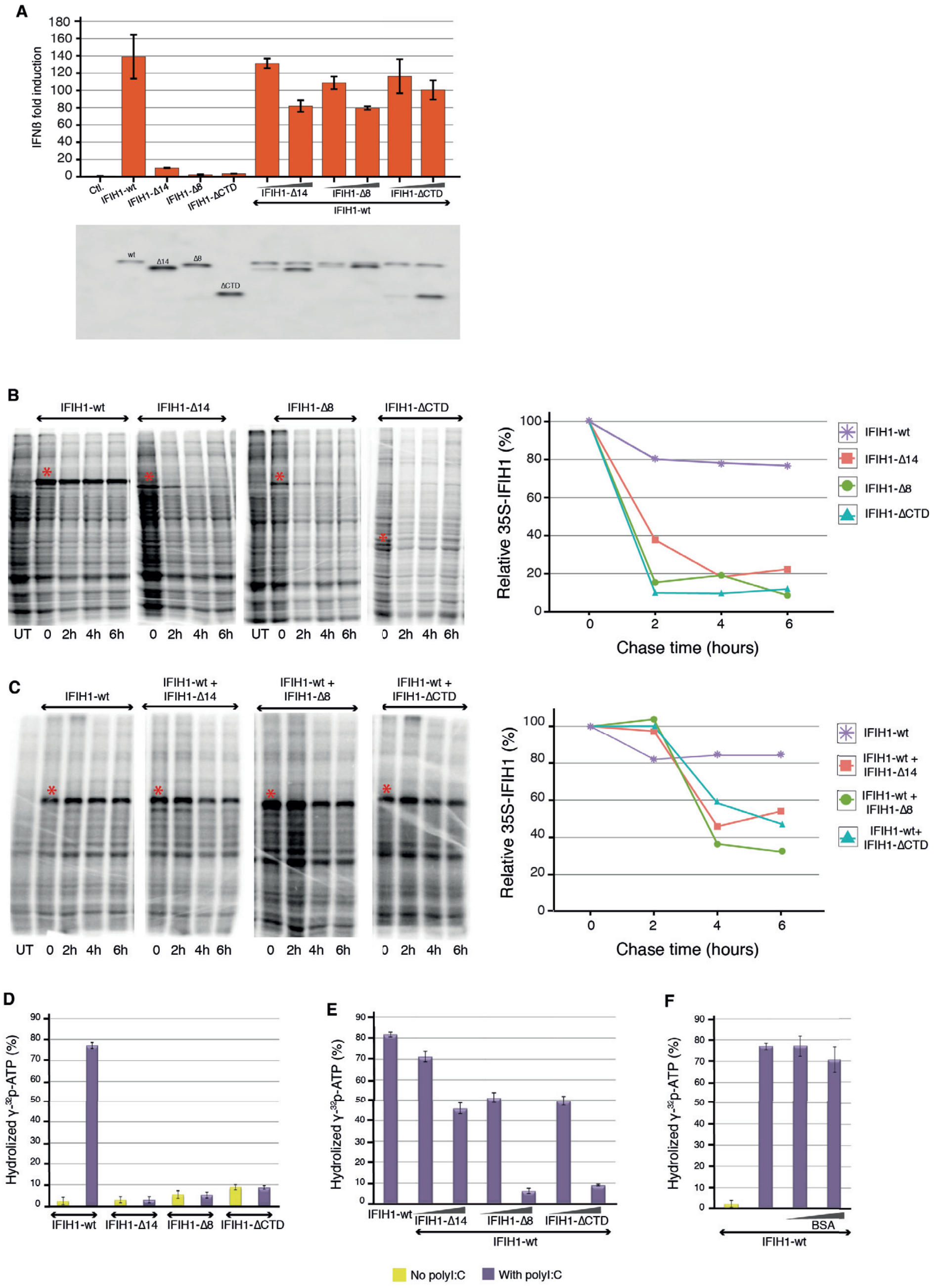

Figure 3.2: Functional characterization of *IFIH1* LoF mutations

(**A**) While transfection with 20ng IFIH1-wt plasmid induce IFNβ in 293T cells, loss-of-function variants in *IFIH1* impair IFNβ induction (40ng or 120ng of each mutant *IFIH1* plasmids, n=3); Co-transfection of 20ng IFIH1-wt plasmid and mutant *IFIH1* plasmids (40ng and 1200ng of *IFIH1* mutant plasmids), co-expression of the alternate isoforms shows a slight interference with IFIH1-wt IFNβ induction (n=3); (**B-C)** Protein stability followed by pulse chase in 293T cells expressing IFIH1-wt, IFIH1-Δ14, IFIH1-Δ8 or IFIH1-ΔCTD; IFIH1-wt has a longer half-life than the alternate IFIH1 isoforms, and the stability of IFIH1-wt is reduced upon co-expression with the alternate isoforms (2ug plasmid expressing wt protein, 2ug of each mutant plasmid, n=1); (**D-E**) RNA-induced ATPase activity of purified IFIH1-wt and alternate IFIH1 isoforms; all three alternate isoforms lack ATPase activity with (purple) or without (yellow) polyI:C stimulation, and IFIH1-wt ATPase activity is reduced upon co-incubation with either of the isoforms in a dose dependent manner (300ng wt protein, 300ng or 600ng of each alternate isoform, 10ng polyI:C, n=2); (**F**) Bovine serum albumin (BSA) does not interfere with ATPase activity of wt protein (300ng wt protein, 300ng or 600ng of BSA, 10ng polyI:C, n=2). Data are represented as mean ± SD. wt: wild-type. polyI:C: polyinosinic:polycytidylic acid.

3.3.10 Role of IFIH1 in controlling respiratory viruses

Viral testing of respiratory samples showed that six of the patients harboring *IFIH1* LoF alleles were infected with RSV and two with human rhinovirus (HRV). To study the effect of IFIH1 on RSV and HRV replication, we used Huh7.5 cells, which lack endogenous expression of IFIH1 and express a mutated, inactive form of RIG-I, and thus are completely unreactive to the RNA pathogen-associated molecular patterns that normally activate these pathways (*11*). The cells were transduced with an IFIH1-expressing lentiviral vector. We observed a much higher level of viral replication in native than in IFIH1-transduced Huh7.5 cells upon infection with HRV-B14, HRV-A16 (Figure 3A-B). The role of IFIH1 in HRV restriction was further demonstrated by ^{35}S labelling of infected cells, which showed a stronger shutoff of cellular protein synthesis in native than in IFIH1-transduced Huh7.5 cells, due to higher replication of HRV-B14 in the absence of IFIH1 (Figure 3C). We observed a much higher level of viral replication in native than in IFIH1-transduced Huh7.5 cells upon RSV infection (Figure 3D-F). We also measured RSV replication in mouse embryonic fibroblasts (MEFs), *ifih1(+/+),* and in IFIH1-knock out MEFs, *ifih1(-/-),* and obtained similar results (Figure 3G-I). Together, these results affirm the central role of IFIH1 in innate immune recognition of RSV and HRV (*12*, *13*). Therefore, LoF variants in *IFIH1* can be reasonably expected to increase susceptibility to these viruses.

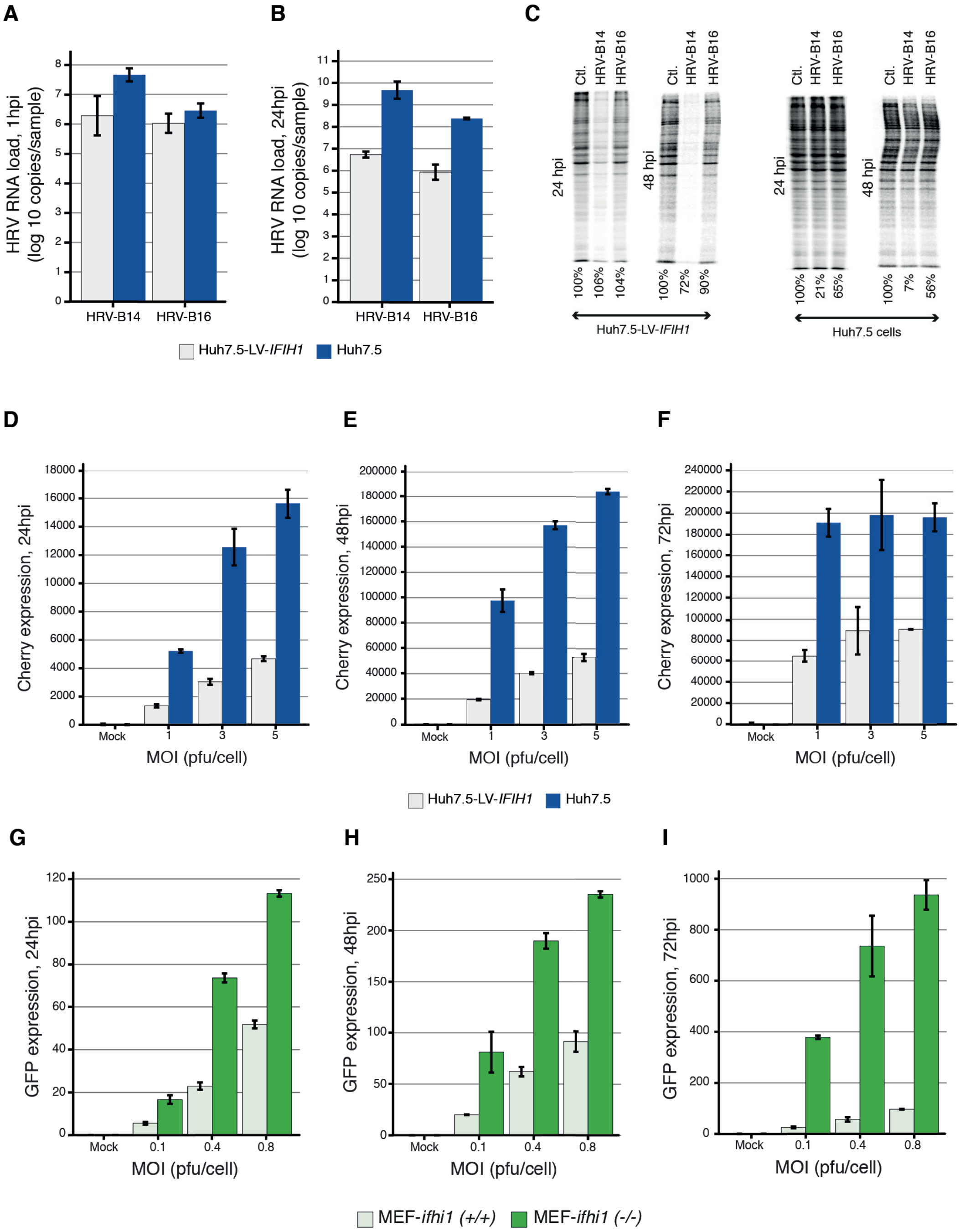

Figure 3.3: Role of IFIH1 in controlling HRV and RSV infections

(A-B) Real time PCR of HRV-B14 (n=6) and HRV-B16 (n=4) RNA in Huh7.5 and huh7.5-LV-*IFIH1* transduced cells 1 and 24 hpi show that HRV-B14 and HRV-B16 replicates more efficiently in the absence of IFIH1; (C) ^{35}S labeling of cellular proteins at 24 and 48 hpi with both HRV-B14, HRV-B16 showing a much stronger shutoff of cellular protein synthesis in huh7.5 cells than in huh7.5-LV-*IFIH1* transduced cells, due to higher viral replication; (D-F) FACS analyses of mCherry-tagged RSV in Huh7.5 cells and Huh7.5-LV-*IFIH1* transduced cells at 24, 48 and 72 hpi show that RSV replicates more efficiently in the absence of IFIH1, (n=2). (H-J) FACS analyses of GFP-tagged RSV in MEF-*ifih1*(-/-) and MEF-LV-*IFIH1* transfected cells at 24, 48 and 72 hpi show that RSV replicates more efficiently in the absence of IFIH1, (n=2). Data are represented as mean ± SD. GFP: green florescent protein, MOI: multiplicity of infection, hpi: hours post infection.

3.4 Conclusion and discussion

It has been shown previously that rare human genetic variants can cause pathogen specific primary immunodeficiencies or confer predisposition to a narrow range of infections in otherwise healthy individuals (*19–26*). We hypothesized that extreme susceptibility to common viral respiratory infection – a rare, potentially lethal phenotype – could also reflect an underlying immunodeficiency. Using an unbiased exome-wide approach in carefully selected study participants, we identified a rare monogenic defect predisposing to severe clinical presentations of RSV and enterovirus/rhinovirus infections. Three deleterious variants were observed in *IFIH1*, which encodes a cytoplasmic receptor critical for viral RNA sensing. The IFIH1 protein binds to dsRNA in an ATP-dependent manner, and initiates a signalling cascade that leads to type 1 interferon production, a process that is disrupted in the presence of any of the LoF variants found in our study population. Interestingly, all three have been shown to be protective against type 1 diabetes (*27*). Common *IFIH1* variants have also been associated with diabetes and other autoimmune diseases (*28–30*), as well as with hepatitis C virus (HCV) clearance (*31*). Additionally, rare gain-of-function mutations in *IFIH1* dramatically upregulate type I interferon production, resulting in Aicardi-Goutières syndrome or Singleton-Merten syndrome (*32–34*). Considered together, the results from all these human genetic studies underscore the pivotal role of innate immune activation in the intricate balance between host defense, inflammation and autoimmunity.

LoF variants in *IFIH1* were only found in a minority of the 120 children enrolled in our study, suggesting that other genetic or non-genetic risk factors remain to be discovered. Genetic heterogeneity is likely to be the rule rather than the exception in extreme presentations of infectious diseases, and larger sample sizes will be required to identify exceedingly rare causal alleles. Whole genome sequencing will also be needed to obtain a more complete coverage of exonic regions (*35*), and to explore non-coding and large-scale structural variation.

The elucidation of the human genetic basis of severe infection provides mechanistic insight into pathogenesis and innate immune response. An immediate practical implication is the possibility to develop diagnostic assays to identify susceptible individuals that could benefit from specific preventive and interventional measures. By highlighting the genes and pathways that play an essential role in host-pathogen interaction, genetic discovery in individuals with extreme phenotypes also provides the opportunity to design new therapeutic strategies that could be useful for the vast majority of patients with milder clinical presentation.

3.5 References

1. Nair, H. *et al.* Global and regional burden of hospital admissions for severe acute lower respiratory infections in young children in 2010: a systematic analysis. *The Lancet* **381,** (2013).
2. Liu, L. *et al.* Global, regional, and national causes of child mortality in 2000-13, with projections to inform post-2015 priorities: an updated systematic analysis. *Lancet* **385,** 430–40 (2015).
3. Jain, S. *et al.* Community-acquired pneumonia requiring hospitalization among U.S. children. *N. Engl. J. Med.* **372,** 835–45 (2015).
4. Debiaggi, M., Canducci, F., Ceresola, E. R. & Clementi, M. The role of infections and coinfections with newly identified and emerging respiratory viruses in children. *Virology journal* **9,** 247 (2012).
5. Rodríguez, D. A. *et al.* Predictors of severity and mortality in children hospitalized with respiratory syncytial virus infection in a tropical region. *Pediatric pulmonology* **49,** 269–76 (2014).
6. Hall, CB, Weinberg, GA & Iwane, MK. The burden of respiratory syncytial virus infection in young children. *& England Journal of &* (2009). doi:10.1056/NEJMoa0804877
7. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–60 (2009).
8. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20,** 1297–303 (2010).
9. Auwera, Carneiro & Hartl. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. (2013). doi:10.1002/0471250953.bi1110s43
10. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature genetics* **46,** 912–8 (2014).
11. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–9 (2009).
12. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6,** 80–92 (2012).
13. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
14. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–74 (2012).
15. King & Goodbourn. The beta-interferon promoter responds to priming through multiple independent regulatory elements. *The Journal of biological chemistry* **269,** 30609–15 (1994).
16. Hallak, L. K., Collins, P. L., Knudson, W. & Peeples, M. E. Iduronic acid-containing glycosaminoglycans on target cells are required for efficient respiratory syncytial virus infection. *Virology* **271,** 264–75 (2000).
17. Rameix-Welti, M.-A. A. *et al.* Visualizing the replication of respiratory syncytial virus in cells and in living mice. *Nat Commun* **5,** 5104 (2014).

18. Ostertag, D., Hoblitzell-Ostertag, T. M. & Perrault, J. Overproduction of double-stranded RNA in vesicular stomatitis virus-infected cells activates a constitutive cell-type-specific antiviral response. *J. Virol.* **81,** 503–13 (2007).

19. Hausmann, S., Marq, J.-B. B., Tapparel, C., Kolakofsky, D. & Garcin, D. RIG-I and dsRNA-induced IFNbeta activation. *PLoS ONE* **3,** e3965 (2008).

20. Kato, H. *et al.* Differential roles of MDA5 and RIG-I helicases in the recognition of RNA viruses. *Nature* **441,** 101–5 (2006).

21. Gitlin, L. *et al.* Essential role of mda-5 in type I IFN responses to polyriboinosinic:polyribocytidylic acid and encephalomyocarditis picornavirus. *Proc. Natl. Acad. Sci. U.S.A.* **103,** 8459–64 (2006).

22. Wang, Q. *et al.* MDA5 and TLR3 initiate pro-inflammatory signaling pathways leading to rhinovirus-induced airways inflammation and hyperresponsiveness. *PLoS Pathog.* **7,** e1002070 (2011).

23. Barral, P. M. *et al.* MDA-5 is cleaved in poliovirus-infected cells. *J. Virol.* **81,** 3677–84 (2007).

24. Wang, J. P. *et al.* MDA5 and MAVS mediate type I interferon responses to coxsackie B virus. *J. Virol.* **84,** 254–60 (2010).

25. Jin, Y.-H. H. *et al.* Melanoma differentiation-associated gene 5 is critical for protection against Theiler's virus-induced demyelinating disease. *J. Virol.* **86,** 1531–43 (2012).

26. Slater, L. *et al.* Co-ordinated role of TLR3, RIG-I and MDA5 in the innate response to rhinovirus in bronchial epithelium. *PLoS Pathog.* **6,** e1001178 (2010).

27. Feng, Q., Langereis, M. A. & van Kuppeveld, F. J. Induction and suppression of innate antiviral responses by picornaviruses. *Cytokine Growth Factor Rev.* **25,** 577–85 (2014).

28. Siu, K.-L. L. *et al.* Middle east respiratory syndrome coronavirus 4a protein is a double-stranded RNA-binding protein that suppresses PACT-induced activation of RIG-I and MDA5 in the innate antiviral response. *J. Virol.* **88,** 4866–76 (2014).

29. Cao, X. *et al.* MDA5 plays a critical role in interferon response during hepatitis C virus infection. *J. Hepatol.* **62,** 771–8 (2015).

30. Gitlin, L. *et al.* Melanoma differentiation-associated gene 5 (MDA5) is involved in the innate immune response to Paramyxoviridae infection in vivo. *PLoS Pathog.* **6,** e1000734 (2010).

31. Grandvaux, N. *et al.* Sustained activation of interferon regulatory factor 3 during infection by paramyxoviruses requires MDA5. *J Innate Immun* **6,** 650–62 (2014).

32. Kim, W.-K. K. *et al.* Deficiency of melanoma differentiation-associated protein 5 results in exacerbated chronic postviral lung inflammation. *Am. J. Respir. Crit. Care Med.* **189,** 437–48 (2014).

33. Baños-Lara, M. D. R. del R., Ghosh, A. & Guerrero-Plata, A. Critical role of MDA5 in the interferon response induced by human metapneumovirus infection in dendritic cells and in vivo. *J. Virol.* **87,** 1242–51 (2013).

34. Shingai, M. *et al.* Differential type I IFN-inducing abilities of wild-type versus vaccine strains of measles virus. *J. Immunol.* **179,** 6123–33 (2007).

35. Loo, Y.-M. M. *et al.* Distinct RIG-I and MDA5 signaling by RNA viruses in innate immunity. *J. Virol.* **82,** 335–45 (2008).

36. Zhong, J. *et al.* Robust hepatitis C virus infection in vitro. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 9294–9 (2005).

37. Casanova, J.-L. L. & Abel, L. The genetic theory of infectious diseases: a brief history and selected illustrations. *Annu Rev Genomics Hum Genet* **14,** 215–43 (2013).

38. Casanova, J.-L. L., Conley, M. E., Seligman, S. J., Abel, L. & Notarangelo, L. D. Guidelines for genetic studies in single patients: lessons from primary immunodeficiencies. *J. Exp. Med.* **211,** 2137–49 (2014).

39. Boisson-Dupuis, S. *et al.* Inborn errors of human STAT1: allelic heterogeneity governs the diversity of immunological and infectious phenotypes. *Curr. Opin. Immunol.* **24,** 364–78 (2012).

40. Zhang, S.-Y. Y. *et al.* TLR3 deficiency in patients with herpes simplex encephalitis. *Science* **317,** 1522–7 (2007).

41. Guo, Y. *et al.* Herpes simplex virus encephalitis in a patient with complete TLR3 deficiency: TLR3 is otherwise redundant in protective immunity. *J. Exp. Med.* **208,** 2083–98 (2011).

42. Pérez de Diego, R. *et al.* Human TRAF3 adaptor molecule deficiency leads to impaired Toll-like receptor 3 response and susceptibility to herpes simplex encephalitis. *Immunity* **33,** 400–11 (2010).

43. Pedergnana, V. *et al.* A major locus on chromosome 3p22 conferring predisposition to human herpesvirus 8 infection. *European journal of human genetics : EJHG* **20,** 690–5 (2012).

44. Ciancanelli, M. J. *et al.* Infectious disease. Life-threatening influenza and impaired interferon amplification in human IRF7 deficiency. *Science* **348,** 448–53 (2015).

45. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J. A. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324,** 387–9 (2009).

46. Smyth, D. *et al.* A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. *Nature genetics* **38,** 617–9 (2006).

47. Yang, H. *et al.* IFIH1 gene polymorphisms in type 1 diabetes: genetic association analysis and genotype-phenotype correlation in Chinese Han population. *Autoimmunity* **45,** 226–32 (2012).

48. Cen, H. *et al.* Association of IFIH1 rs1990760 polymorphism with susceptibility to autoimmune diseases: a meta-analysis. *Autoimmunity* **46,** 455–62 (2013).

49. Hoffmann, F. *et al.* Polymorphisms in melanoma differentiation-associated gene 5 link protein function to clearance of hepatitis C virus. *Hepatology (Baltimore, Md.)* **61,** 460–70 (2015).

50. Rice, G. I. *et al.* Gain-of-function mutations in IFIH1 cause a spectrum of human disease phenotypes associated with upregulated type I interferon signaling. *Nat. Genet.* **46,** 503–9 (2014).

51. Oda, H. *et al.* Aicardi-Goutières syndrome is caused by IFIH1 mutations. *American journal of human genetics* **95,** 121–5 (2014).

52. Rutsch, F. *et al.* A specific IFIH1 gain-of-function mutation causes Singleton-Merten syndrome. *Am. J. Hum. Genet.* **96,** 275–82 (2015).

53. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511,** 344–347 (2014).

Chapter 4 Deep sequencing and *de novo* assembly of two respiratory syncytial virus genomes isolated from previously healthy children with severe infection

4.1 Introduction

Human Respiratory Syncytial Virus (RSV) is the leading cause of lower respiratory tract infection in children worldwide. Between 50 and 70% of newborns get infected with RSV within their first year of life and almost all children suffer from RSV infection before the age of two (*1*, *2*). RSV belongs to the family paramyxoviridae. It is an enveloped, negative-sense RNA virus with a single-stranded genome of approximately 15 kb (*3*). The nonsegmented RSV genome encodes 11 proteins. Small hydrophobic protein (SHpr), fusion glycolprotein (Fpr) and attachment glycoprotein (Gpr) are transmembrane glycoproteins. Fpr and Gpr are the two major glycoproteins on RSV's surface that mediates the attachment and entry of the virus to the host cells. Matrix protein (Mpr) is an internal membrane associated matrix protein; nucleoprotein (Npr), phosphoprotein (Ppr), matrix M2-1 protein (M2-1pr) and RNA polymerase (Lpr) are components of the nucleocapsid/polymerase complex; nonstructural protein 1 and 2 (NS1 and NS2) are presumptive nonstructural proteins that act against the host immune system; and matrix M2-2 protein (M2-2pr) is a low abundance protein that is suspected to play a regulatory role in RNA transcription and translation (*4*, *5*). There are two main RSV strains: RSVA and RSVB. At the serological level, the two strains are characterized by different antibody responses against Fpr and Gpr (*6*). Each strain is further subdivided into different genotypes according to phylogenetic analyses of specific genomic regions.

In the vast majority of cases, RSV infection manifests itself as a normal cold, with spontaneous resolution of respiratory symptoms after a few days (*7*). Beside socio-economic and environmental factors (e.g. mal-nutrition and smoking parents), a number of risk factors have been described for severe susceptibility to respiratory viruses including male gender, preterm birth, chronic diseases and immune suppression (*8–11*). However, approximately 1 out of 1000 children without any known risk factor or co-morbidity requires intensive care support due to life-threatening manifestations of RSV infection. Our recent work showed that rare protein truncating mutations in the *IFIH1* gene lead to extreme susceptibility to RSV in a subset of otherwise healthy children (Asgari et al., submitted). A number of earlier studies focused on RSV genetic diversity and suggested that the severity of RSV infection depends in part on viral strain or in a few cases on variations of large effects (deletions and duplications) in the viral genome (*12*–*14*).

Here, we sequence and reconstruct the complete genomes of two RSV strains found in children with very severe RSV disease, who have also been fully characterized at the exome and blood transcriptome levels. We further compare them with 53 recently sequenced RSV genomes isolated mainly from patients with mild symptoms (*15*). This study aims at establishing a framework to explore the genetic diversity of RSV strains recovered from patients with unusually severe disease, in an attempt to determine whether viral genetic factors are partly responsible for the most severe clinical presentations.

4.2 Material and methods

4.2.1 Subject recruitment and specimen collection

Study participants were recruited at the Royal Children's Hospital (RCH) and Mater Children's Hospital (MCH) in Brisbane, Australia, in 2013. The study was approved by the relevant Ethics Commissions and written informed consent was obtained from parents or legal guardian. A nasopharyngeal aspirate (NPA) was collected from previously healthy children who needed admission to the intensive care unit (ICU) and ventilation support due to severe lower respiratory tract RSV infection. Viral RNA was extracted individually (100ul sample + 300ul PBS) using the NucliSens Easymag© according to the manufacturer's instructions. Respiratory virus screening was performed using the FTD Respiratory pathogens 21 assay on a Viia7 instrument. The viral strain was determined prior to sequencing using PCR with strain specific primers.

4.2.2 Viral RNA sequencing

We sequenced one sample positive for RSVA (PRI-024) and one sample positive for RSVB (PRI-030). The number of viral RNA copies was inferior to the limit of detection of the quantitative PCR assay. Ribosomal RNA (rRNA) was removed using a Ribo-Zero Gold kit according to the manufacturer's protocol. rRNA-depleted specimens were purified on Zymo columns. Libraries were prepared with the low-throughput TruSeq total RNA preparation protocol from Illumina using 15 PCR cycles. Library concentrations were measured with Q-bit. Size distribution of fragments was estimated with 2200 TapeStation. Fragments of 200 to 450 bp were obtained. All specimens were sequenced paired-end using the 100-bp protocol with indexing on a HiSeq 2500 sequencer in pools of two specimens per lane. RNA-seq libraries were loaded at 8 pM.

4.2.3 De novo assembly and analysis

Short reads from each sample were mapped to the human reference genome GRCh37 to identify and remove human sequences. Unmapped paired-end reads were extracted and used for *de novo* assembly using the default options of CLC genomics workbench (https://www.qiagenbioinformatics.com/). To make contigs, the CLC genomics workbench *de novo* assembler used Bruijn graphs, an efficient algorithm allowing fast contig formation. The basic idea is to hash the reads according to a predefined *k*-mer length called word size. In the next step, for a word size of *k* the algorithm will find all the neighbouring words of *k*-1 overlapping ends. The ideal word size is a balance between specificity and sensitivity. A de Bruijn graph is made of each word as a node and each neighbour as an edge. In an ideal situation only one forward neighbour and one backward neighbour is available for each word and long, linear stretches of connected nodes can be made. In many situations, however, many nodes have more than one incoming or outgoing edge. The de Bruijn graph is resolved by following a Eulerian trail in the graph, a trail that visits every edge exactly once. For *de novo* assembly we used the word size of 23 and minimum contig length of 2500 bp. To identify the produced contigs we used the generated contigs as query sequences for NCBI nucleotide BLAST to find the best matches to the query sequences, defined as matches with the lowest E-value and the highest sequence similarity.

4.2.4 Maximum likelihood phylogeny analysis

The newly sequenced and assembled genomes were compared with 53 complete RSV genomes (34 RSVA and 19 RSVB, BioProject PRJNA73049, Table 4.3). These 53 sequences are fully characterized RSV genomes from respiratory samples collected in Mexico, Argentina, Belgium, Italy, Germany, Australia, South Africa, and the USA between 1998-2010. These samples were collected from adult and infant patients with normal / mild clinical presentation of the disease to investigate RSV genomic diversity. Multiple alignments of PRI-024 and PRI-030 genome sequence with these publically available genomes were used for inferring a maximum likelihood tree with 100 bootstraps using CLC genomics workbench.

4.2.5 Protein similarity analysis

PRI-024 and PRI-030 were annotated for ORF using VIGOR (http://jcvi.org/vigor/index.php) and the sequence of each of the 11 RSV proteins was obtained. We annotated the 34 RSVA and 19 RSVB listed in table 4.3 and obtained a consensus sequence for each RSV protein by multiple alignments of these protein sequences. We used dot plots to search for gaps or stretches of low similarity between our samples and the consensus protein sequences from the above-mentioned RSV genomes. To look for single amino acid changes, we used multiple alignments of our samples and the 34 RSVA and 19 RSVB listed in table 4.3.

4.3 Results

4.3.1 Study subjects

Sample PRI-024 was extracted from a 72 day old female and sample PRI-030 was extracted from a 21 day old male, both of recent Western European ancestry. They required ICU admission and non-invasive ventilatory support for respiratory symptoms attributed to RSV bronchiolitis. No co-infection or co-morbidity was reported. Both had a full recovery and their medical history was otherwise unremarkable. A separate analysis of their germline DNA (by exome sequencing) and transcriptome (by RNA sequencing) did not reveal any information that could explain the severity of the presentation (Asgari et al., submitted).

4.3.2 Viral RNA sequencing

For PRI-024, 149,805,670 paired-end reads were obtained. 96.2% of reads had phred scale sequence quality score ≥ 20 (88.5% ≥ 30). 64% of the reads were mapped to the human reference genome. 30% of the reads were paired-end reads that were not mapped to human genome.

For PRI-030, 144,588,304 paired-end reads were obtained. 96.5% of reads had phred scale sequence quality score ≥ 20 (86% ≥ 30). 47.2% of the reads were mapped to human reference genome. 49.6% of the reads were paired-end reads that were not mapped to human genome.

4.3.3 *de novo* assembly and analysis

For *de novo* assembly we used all paired-end reads that were not mapped to human reference genome (the reads that were mapped to human reference genome and all the single reads were removed). For PRI-024, 45,018,680 paired-end reads were used. N50 was 3,247 and the longest contig was 15,219 bp, resulting from the assembly of 5,388,825 reads (11.97% of reads) (Figure 4.1 A and C). For PRI-030 71,837,152 paired-end reads were used. N50 was 3,281 and the longest contig was 15,321 bp, resulting from the assembly of 4,988,115 reads (6.94% of reads) (Figure 4.1 B and D). In both cases, the longest contig had a high coverage (34000x on average) and matched the expected length of the RSV genome.

To identify the reconstructed contigs we used them as query sequences for nucleotide BLAST against the NCBI nucleotide database. For PRI-024, 72 contigs with length > 2500bp was generated. For the longest PRI-024 contig, a human RSVA was the best match (accession: KJ627365). For PRI-030, 44 contigs with length > 2500bp was generated. For PRI-030 the longest PRI-030 longest contig a human RSVB was the best match (accession: JX576745). The BLAST analysis of these contigs allowed the identification of additional viral and bacterial species in the samples. Most had no known correlation with respiratory infections, but a single contig in PRI-030 showed high similarity to Moraxella catarrhalis strain 25240. (Table 4.1 and 4.2)

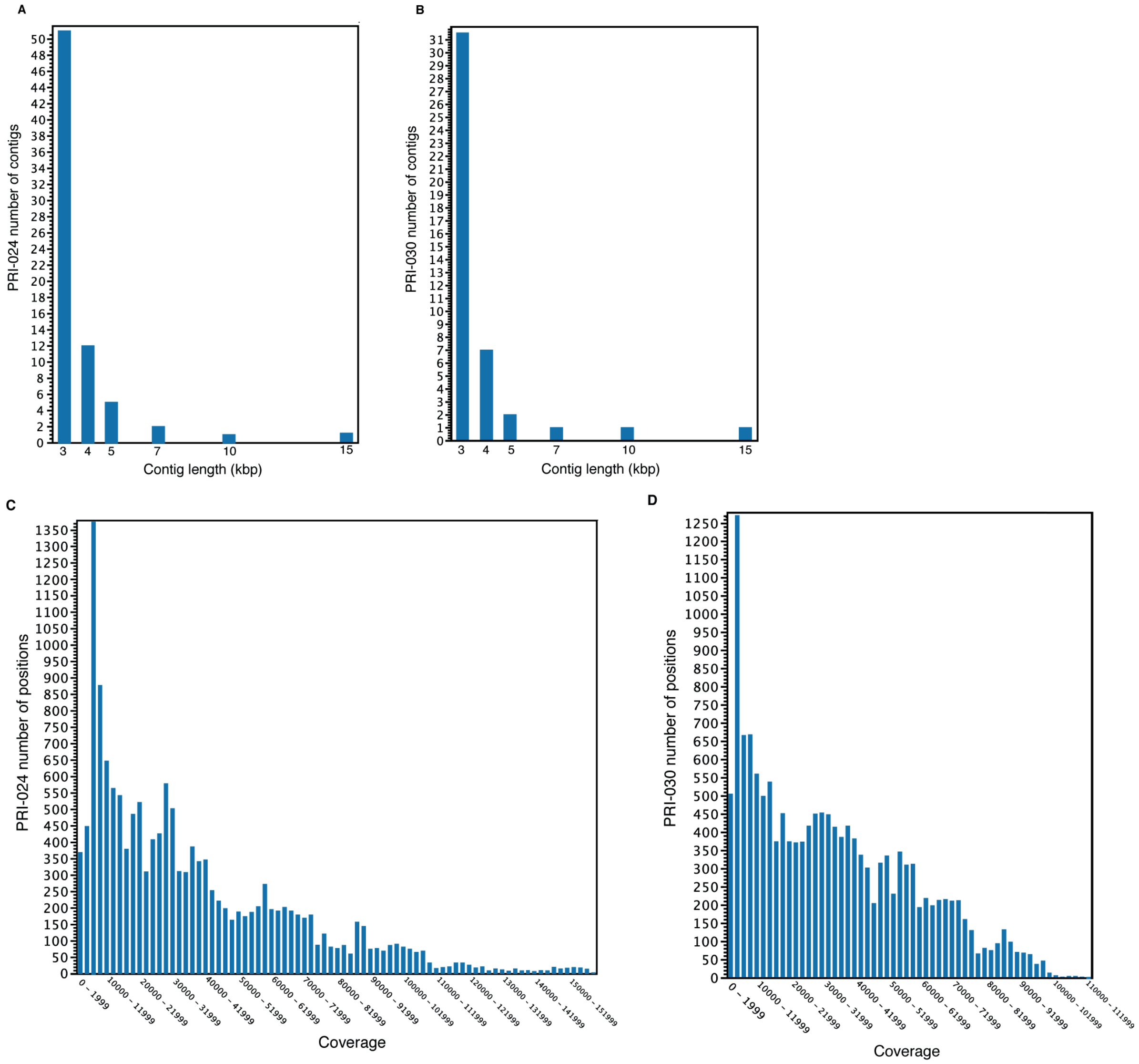

Figure 4.1: *de novo* assembly of RSV genome

A & B) de novo assembly contig length distribution; in each assembly the longest contig was around 15,000 bp, which is the expected length of the RSV genome. C & D) Coverage distribution for the longest contig from each assembly. Both contigs were deeply sequenced and fully covered.

Table 4.1: BLAST analysis for PRI-024 *de novo* assembled contigs

BLAST analysis of 72 contigs generated from *de novo* assembly of PRI-024 sample showed that the best matching sequence to the longest contig is a RSVA genome, while the remaining contigs map to mostly yeast genomes.

Query	Lowest E-value	Accession for lowest E-value	Description for lowest E-value
PRI-024_contig_1	0	KJ627365	Human respiratory syncytial virus strain RSVA
PRI-024_contig_2	0	HF558455	Rhodotorula taiwanensis RS1
PRI-024_contig_3	0	LK052953	Rhodosporidium toruloides strain CECT1137
PRI-024_contig_4	0	HF558455	Rhodotorula taiwanensis RS1
PRI-024_contig_5	0	HF558455	Rhodotorula taiwanensis RS1
PRI-024_contig_6	0	XM_002548769	Candida tropicalis MYA-3404 C-1-tetrahydrofolate synthase
PRI-024_contig_7	0	XM_002616278	Clavispora lusitaniae ATCC 42720 conserved hypothetical protein
PRI-024_contig_8	0	XM_001382415	Scheffersomyces stipitis CBS 6054 plasma membrane H+ATPase
PRI-024_contig_9	0	XM_007385036	Punctularia strigosozonata HHB-11173 SS5 hypothetical protein, (PUNSTDRAFT_104183)
PRI-024_contig_10	0	LK052965	Rhodosporidium toruloides strain CECT1137
PRI-024_contig_11	0	KR055655	Pseudogymnoascus pannorum strain NN050741 mitochondrion
PRI-024_contig_12	6.5E-169	LK052940	Rhodosporidium toruloides strain CECT1137, genomic scaffold
PRI-024_contig_13	0	HF558455	Rhodotorula taiwanensis RS1, complete mitochondrial genome
PRI-024_contig_14	0	XM_007917050	Togninia minima UCRPA7 putative plasma membrane atpase protein mRNA

PRI-024_contig_15	0	LK052936	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S01
PRI-024_contig_16	3.7E-69	XM_001483899	Meyerozyma guilliermondii ATCC 6260 hypothetical protein (PGUG_03330) partial RNA
PRI-024_contig_17	9.6E-83	CP003834	Cryptococcus neoformans var. grubii H99 mitochondrion, complete genome
PRI-024_contig_18	0	XM_002548335	Candida tropicalis MYA-3404 elongation factor 3, mRNA
PRI-024_contig_19	5.3E-143	LK052943	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S08
PRI-024_contig_20	0	FQ311441	Sporisorium reilianum SRZ2 chromosome 2 complete DNA sequence
PRI-024_contig_21	6E-161	XM_009547565	Heterobasidion irregulare TC 32-1 oligopeptide transporter mRNA
PRI-024_contig_22	0	XM_014321648	Trichosporon asahii var. asahii CBS 2479 delta-9 fatty acid desaturase partial RNA
PRI-024_contig_23	0	LK052949	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S14
PRI-024_contig_24	0	LK052954	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S19
PRI-024_contig_25	0	CP000502	Scheffersomyces stipitis CBS 6054 chromosome 8, complete sequence
PRI-024_contig_26	0	HF558455	Rhodotorula taiwanensis RS1, complete mitochondrial genome
PRI-024_contig_27	0	LK052942	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S07
PRI-024_contig_28	0	LK052939	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S04
PRI-024_contig_29	0	XM_013568353	Aureobasidium namibiae CBS 147.97 hypothetical protein partial mRNA
PRI-024_contig_30	0	LK052951	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S16
PRI-024_contig_31	0	LK052965	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S30

PRI-024_contig_32	0	LK052940	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S05
PRI-024_contig_33	0	LK052945	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S10
PRI-024_contig_34	0	LK052940	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S05
PRI-024_contig_35	0	KC616429	Scheffersomyces coipomoensis isolate NRRL_Y-17651 endogenous virus ScV, putative proteins and tryptophan-2,3-dioxygenase genes, complete cds
PRI-024_contig_36	0	LK052940	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S05
PRI-024_contig_37	0	CP012999	Enterobacter sp. E20, complete genome
PRI-024_contig_38	0	XM_007374069	Spathaspora passalidarum NRRL Y-27907 hypothetical protein, (SPAPADRAFT_134536), partial mRNA
PRI-024_contig_39	0	XM_002545697	Candida tropicalis MYA-3404 phosphoribosylformylglycinamidine synthase, mRNA
PRI-024_contig_40	0	XM_007372211	Spathaspora passalidarum NRRL Y-27907 beta-1,3-glucan synthase, (SPAPADRAFT_145419), mRNA
PRI-024_contig_41	0	LK052948	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S13
PRI-024_contig_42	0	XM_014800750	Pseudozyma antarctica carbamoyl-phosphate synthase partial mRNA
PRI-024_contig_43	0	XM_006965283	Trichoderma reesei QM6a predicted protein (TRIREDRAFT_121820), mRNA
PRI-024_contig_44	5.8E-67	LK052938	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S03
PRI-024_contig_45	0	XM_014803277	Pseudozyma antarctica family 3 glycosyltransferase partial mRNA
PRI-024_contig_46	1.4E-131	LK052942	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S07
PRI-024_contig_47	0	XM_007374531	Spathaspora passalidarum NRRL Y-27907 hypothetical protein, (SPAPADRAFT_66085), mRNA

PRI-024_contig_48	0	XM_014797905	Pseudozyma antarctica Na/K ATPase alpha 1 subunit partial mRNA
PRI-024_contig_49	0	FQ311435	Sporisorium reilianum SRZ2 chromosome 14 complete DNA sequence
PRI-024_contig_50	0	XM_013566968	Aureobasidium namibiae CBS 147.97 hypothetical protein partial mRNA
PRI-024_contig_51	0	XM_007373532	Spathaspora passalidarum NRRL Y-27907 hypothetical protein, (SPAPADRAFT_49095), partial mRNA
PRI-024_contig_52	0	LK052939	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S04
PRI-024_contig_53	0	LK052936	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S01
PRI-024_contig_54	0	LK052949	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S14
PRI-024_contig_55	0	AB920149	Cryptococcus tepidarius gene for plasma membrane H(+)-ATPase 1, complete, sequence, strain: JCM 11965
PRI-024_contig_56	0	LK052939	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S04
PRI-024_contig_57	4.9E-68	LK052943	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S08
PRI-024_contig_58	0	LK052941	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S06
PRI-024_contig_59	0	LK052886	Cyberlindnera fabianii strain YJS4271 genomic scaffold, scaffold CYFA0S01
PRI-024_contig_60	4.6E-05	CP012620	Oryza sativa Indica Group cultivar RP Bio-226 chromosome 12 sequence
PRI-024_contig_61	0	XM_007880608	Pseudozyma flocculosa PF-1 hypothetical protein partial mRNA
PRI-024_contig_62	0.32	XM_013458970	Exophiala xenobiotica dihydroxy-acid dehydratase mRNA
PRI-024_contig_63	0	XM_001383177	Scheffersomyces stipitis CBS 6054 glycine cleavage system protein partial mRNA
PRI-024_contig_64	0	LK052940	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S05

PRI-024_contig_65	0	LK052945	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S10
PRI-024_contig_66	0	LK052946	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S11
PRI-024_contig_67	0	XM_014318297	Grosmannia clavigera kw1407 elongation factor 3 partial mRNA
PRI-024_contig_68	0	LK052938	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S03
PRI-024_contig_69	9.7E-78	XM_007372681	Spathaspora passalidarum NRRL Y-27907 ran binding protein, (SPAPADRAFT_53613), partial mRNA
PRI-024_contig_70	0	LK052939	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S04
PRI-024_contig_71	0	LK052953	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S18
PRI-024_contig_72	8.8E-65	LK052944	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold, RHTO0S09

Table 4.2: BLAST analysis for PRI-030 *de novo* assembled contigs

BLAST analysis of 44 contigs generated from *de novo* assembly of PRI-030 sample showed that the best matching sequence to the longest contig is an RSVB genome, while the remaining contigs map to mostly commensal genomes.

Query	Lowest E-value	Accession for lowest E-value	Description for lowest E-value
PRI-030_contig_1	0	JX576745	Human respiratory syncytial virus strain 08-045952 from Netherlands, complete genome
PRI-030_contig_2	0	CP007457	Bifidobacterium pseudolongum PV8-2, complete genome
PRI-030_contig_3	0	DQ831892	Fusarium oxysporum f. sp. lactucae isolate FK09701 28S ribosomal RNA gene, partial sequence; 28S- 18S ribosomal RNA intergenic spacer, complete sequence; and 18S ribosomal RNA gene, partial
PRI-030_contig_4	0	XM_001382817	Scheffersomyces stipitis CBS 6054 Elongation factor partial mRNA
PRI-030_contig_5	0	AM905426	Bacterium Ellin514 23S rRNA gene, isolate Ellin514
PRI-030_contig_6	0	CP000094	Pseudomonas fluorescens Pf0-1, complete genome

PRI-030_contig_7	1.8E-09	AM087550	Ophiostoma mitovirus 3b RdRp gene for RNA-dependent RNA polymerase, genomic RNA
PRI-030_contig_8	0	CP012752	Kibdelosporangium phytohabitans strain KLBMP1111, complete genome
PRI-030_contig_9	0	HF558455	Rhodotorula taiwanensis RS1, complete mitochondrial genome
PRI-030_contig_10	0	AY874423	Fusarium oxysporum voucher VPRI 19292 mitochondrion, partial genome
PRI-030_contig_11	0	XM_012193664	Cryptococcus neoformans var. grubii H99 elongation factor 3 mRNA
PRI-030_contig_12	0	CP001630	Actinosynnema mirum DSM 43827, complete genome
PRI-030_contig_13	0	X73672	T.papilionaceus mitochondrion gene for large subunit rRNA, exons 1 and 2
PRI-030_contig_14	0	HF679031	Fusarium fujikuroi IMI 58289 draft genome, chromosome FFUJ_chr09
PRI-030_contig_15	0	XM_013570087	Aureobasidium namibiae CBS 147.97 elongation factor 3 partial mRNA
PRI-030_contig_16	0	XM_013575636	Aureobasidium namibiae CBS 147.97 hypothetical protein partial mRNA
PRI-030_contig_17	0	HF679024	Fusarium fujikuroi IMI 58289 draft genome, chromosome FFUJ_chr02
PRI-030_contig_18	0.007	CP001843	Treponema primitia ZAS-2, complete genome
PRI-030_contig_19	0	KJ807813	Bean common mosaic virus isolate PStV-JX014, complete genome
PRI-030_contig_20	0	CP007155	Kutzneria albida DSM 43870, complete genome
PRI-030_contig_21	0	HE804045	Saccharothrix espanaensis DSM 44229 complete genome
PRI-030_contig_22	0	XM_013568353	Aureobasidium namibiae CBS 147.97 hypothetical protein partial mRNA
PRI-030_contig_23	0	HF558455	Rhodotorula taiwanensis RS1, complete mitochondrial genome
PRI-030_contig_24	0	HF558455	Rhodotorula taiwanensis RS1, complete mitochondrial genome
PRI-030_contig_25	4.7E-169	LK052940	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold RHTO0S05
PRI-030_contig_26	1E-164	XM_013586131	Medicago truncatula mu-like prophage flumu protein gp28 partial mRNA
PRI-030_contig_27	0	KT598234	Soybean-associated single stranded RNA virus 2 isolate SaSSRV2-1, partial sequence
PRI-030_contig_28	6.9E-136	CP003834	Cryptococcus neoformans var. grubii H99 mitochondrion, complete genome

PRI-030_contig_29	0	LK052953	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold RHTO0S18
PRI-030_contig_30	0	CP000094	Pseudomonas fluorescens Pf0-1, complete genome
PRI-030_contig_31	0	HF679026	Fusarium fujikuroi IMI 58289 draft genome, chromosome FFUJ_chr04
PRI-030_contig_32	0	LK052944	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold RHTO0S09
PRI-030_contig_33	0.2	XM_010538616	PREDICTED: Tarenaya hassleriana zinc finger protein 2 (LOC104811796), mRNA
PRI-030_contig_34	0	CP003880	Pseudomonas sp. UW4, complete genome
PRI-030_contig_35	0	HF679025	Fusarium fujikuroi IMI 58289 draft genome, chromosome FFUJ_chr03
PRI-030_contig_36	1.3E-36	CP007410	Pseudomonas brassicacearum strain DF41, complete genome
PRI-030_contig_37	0	CP009290	Pseudomonas chlororaphis subsp. aurantiaca strain JD37, complete genome
PRI-030_contig_38	0	XM_007006344	Tremella mesenterica DSM 1558 plasma membrane H+ ATPase (TREMEDRAFT_40283), mRNA
PRI-030_contig_39	0	LK052942	Rhodosporidium toruloides strain CECT1137, genomic scaffold, scaffold RHTO0S07
PRI-030_contig_40	8E-104	CP003170	Actinoplanes sp. SE50/110, complete genome
PRI-030_contig_41	0	CP012999	Enterobacter sp. E20, complete genome
PRI-030_contig_42	0	CP000094	Pseudomonas fluorescens Pf0-1, complete genome
PRI-030_contig_43	0	HE804045	Saccharothrix espanaensis DSM 44229 complete genome
PRI-030_contig_44	0	CP008804	Moraxella catarrhalis strain 25240, complete genome

4.3.4 Maximum likelihood tree generation

After multiple sequence alignment of PRI-024 with RSVA sequences and of PRI-030 with RSVB sequences listed in table 4.3 we used a maximum likelihood approach to infer the phylogenic tree for both RSV genomes. Each tree includes viruses that have already been assigned to genotypes GA5 and GA2 for RSVA and BA for RSVB. PRI-024 clustered with GA2, which is one of the predominant circulating RSVA genotypes. PRI-030 clustered with BA genotype, the dominant genotype for RSVB worldwide (*15*) (Figure 4.2).

Table 4.3: Pairwise alignment results

34 RSVA and 19 RSVB full genomes were downloaded form GenBank. Our two *de novo* assembled genomes were highly similar (on average 96.36% for PRI-024 and 97.94% for PRI-030) in their sequence to the previously sequenced RSV genomes.

Sample	Strain	Year collected	Genome size	Identity with de novo assembled RSV (%)
KF530258	RSVA	2007	15123	98.28
KF530260	RSVA	2005	14980	94.19
KF530261	RSVA	2008	15100	98.23
KF530263	RSVA	2007	15118	98.08
KF530264	RSVA	2008	15039	98.02
KF530267	RSVA	2007	15078	96.30
KF530268	RSVA	2007	15128	94.81
KF530269	RSVA	2009	15049	97.89
KF826816	RSVA	2005	14948	92.87
KF826817	RSVA	2009	14744	93.43
KF826821	RSVA	2007	15177	98.39
KF826823	RSVA	1998	15197	95.34
KF826824	RSVA	1998	15200	95.23
KF826826	RSVA	2004	15197	95.08
KF826827	RSVA	2004	15209	95.10
KF826828	RSVA	2004	15191	95.13
KF826830	RSVA	2009	15186	98.39
KF826831	RSVA	2009	15195	98.35
KF826832	RSVA	2009	15176	95.04
KF826833	RSVA	2009	15129	98.50
KF826836	RSVA	2006	15165	94.87
KF826837	RSVA	2006	15194	95.02
KF826838	RSVA	2006	15189	98.45
KF826840	RSVA	2007	15106	97.84
KF826841	RSVA	2007	15194	95.05
KF826846	RSVA	2008	15142	94.90
KF826847	RSVA	2007	15179	95.36
KF826848	RSVA	2007	15190	97.36
KF826849	RSVA	2010	15182	98.52
KF826850	RSVA	2008	15193	95.01
KF826852	RSVA	2007	15167	95.04
KF826854	RSVA	2009	15192	94.89
KF826855	RSVA	2009	15163	98.74
KF826856	RSVA	2009	15188	98.59
KF530259	RSVB	2006	15157	98.45

KF530262	RSVB	2009	15045	97.60
KF530266	RSVB	2008	15064	97.78
KF826820	RSVB	2009	15181	94.66
KF826822	RSVB	2007	15269	98.80
KF826825	RSVB	2004	15259	98.13
KF826829	RSVB	2005	15215	97.72
KF826834	RSVB	2009	15218	98.31
KF826835	RSVB	2009	15238	96.43
KF826839	RSVB	2006	15278	98.63
KF826842	RSVB	2007	15269	98.72
KF826843	RSVB	2008	15280	98.47
KF826844	RSVB	2008	15218	98.24
KF826845	RSVB	2008	15147	97.78
KF826851	RSVB	2007	15279	98.79
KF826853	RSVB	2008	15183	96.43
KF826857	RSVB	2009	15278	98.71
KF826858	RSVB	2009	15233	98.73
KF826860	RSVB	2009	15249	98.54

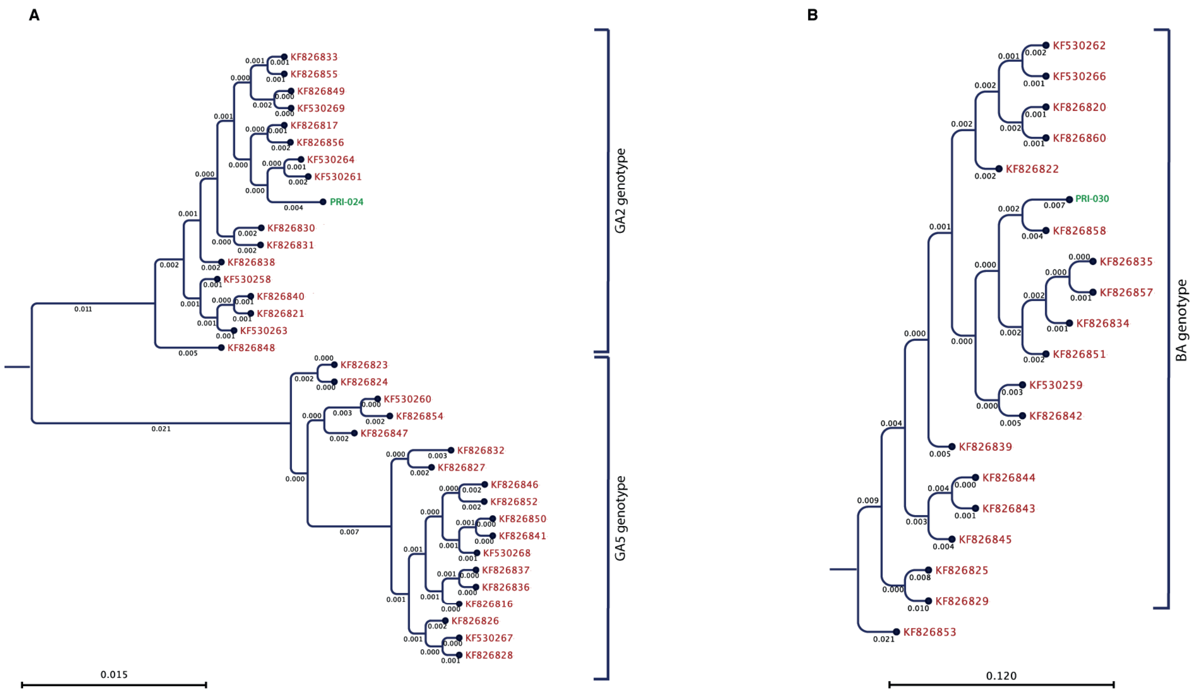

Figure 4.2: Maximum likelihood phylogeny tree

Maximum likelihood phylogeny trees of RSV genome sequences, including PRI-024 (A) and PRI-030 (B) genome sequences. This maximum likelihood phylogeny trees are made using the RSVA and RSVB sequenced in this study plus 34 RSVA and 19 RSVB retrieved from GenBank. Average number of substitution per site is shown on the branches. In panel A, branches shorter than 0.0014 in length are shown as having a length of 0.0014. In panel B branches shorter than 0.126 in length are shown as having a length of 0.126.

4.3.5 Protein similarity analysis

We annotated the ORFs of our two reconstructed RSV genomes and obtained the sequence of the 11 RSV proteins. We compared these protein sequences with consensus sequences derived from RSVA proteins in GA2 clade for PRI-024 (15 genomes) and BA clade for PRI-030 (18 genomes). For both strains all proteins were highly similar to consensus sequence proteins in amino acid composition, amino acid sequence and protein sequence length (Figure 4.3 and 4.4). To get a more detailed picture of RSV protein variation we also performed a multiple alignment of PRI-024 and PRI-030 proteins with other genomes in the same clade and for each protein listed the number of unique amino acids changes. Unique changes were defined as amino acid changes that were observed only once in multiple alignment. Neither PRI-024 nor PRI-030 showed unusually higher variability (Table 4.4).

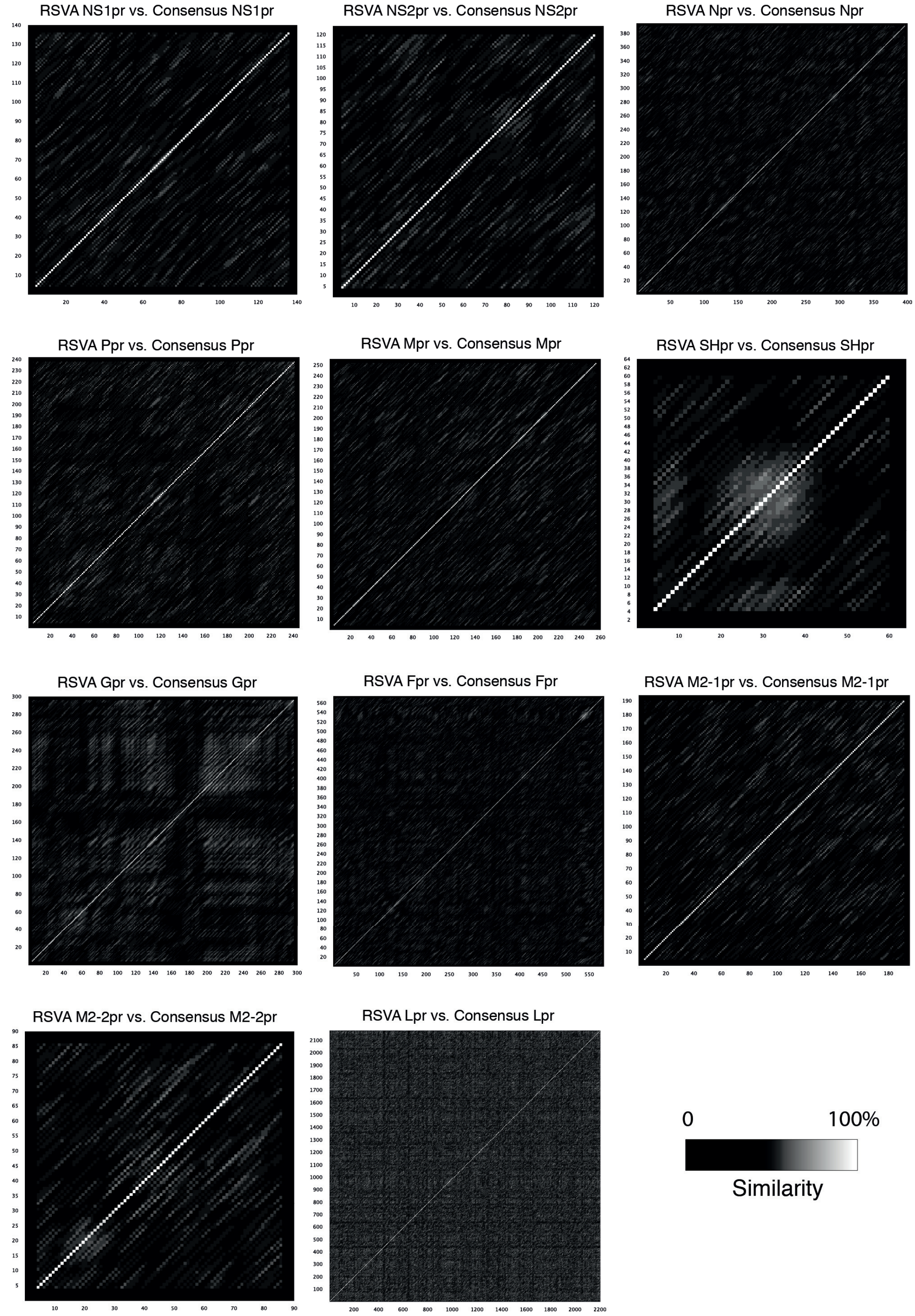

Figure 4.3: PRI-024 protein similarity dot plot

Dot plot showing the level of similarity between PRI-024 proteins translated from the RSVA sequenced in this study and the GA2 clade consensus sequence proteins. The Y axis shows positions in PRI-024 protein sequence and the X axis positions in the consensus protein sequence.

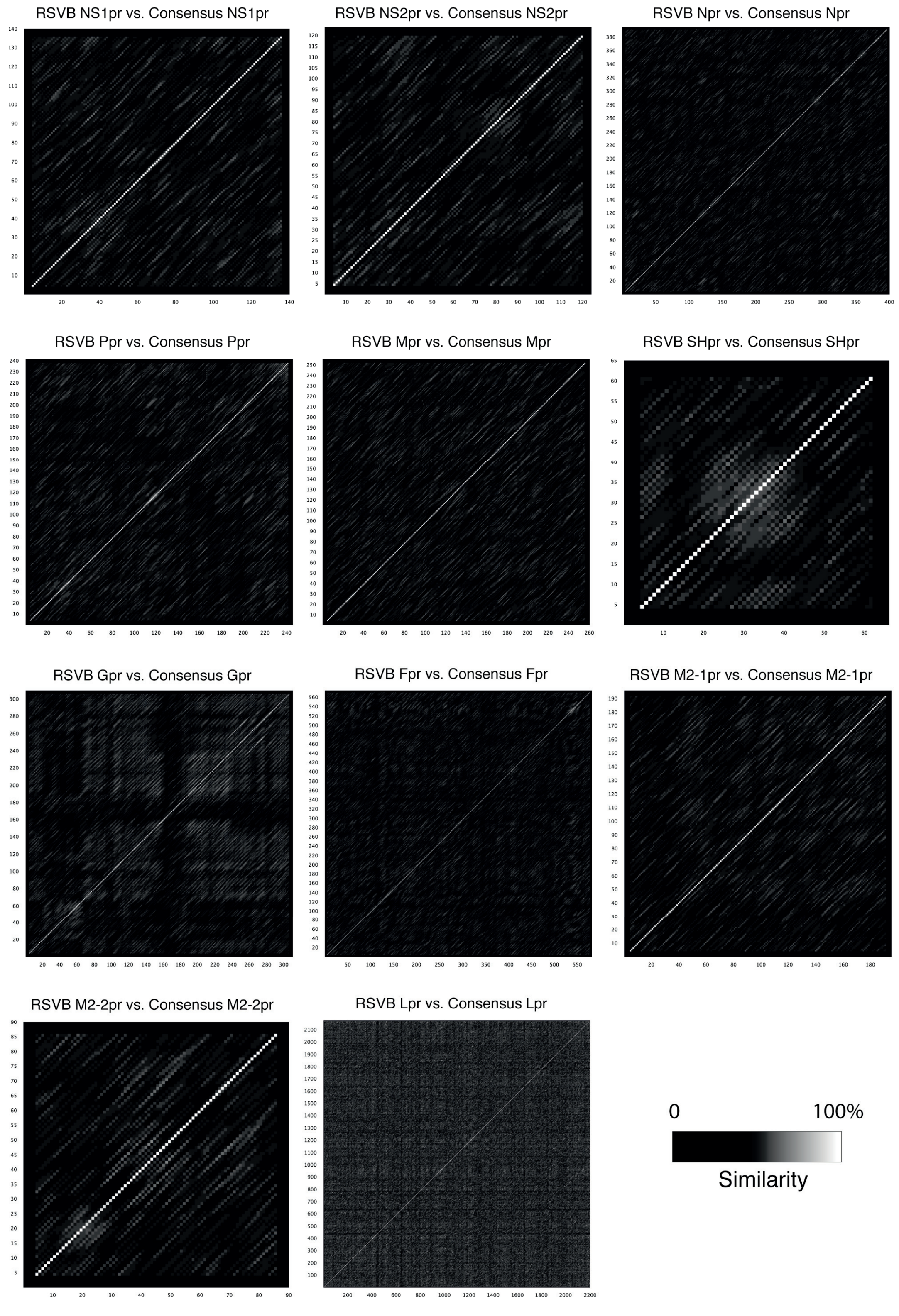

Figure 4.4: PRI-030 protein similarity dot plot

Dot plot showing the level of similarity between PRI-030 proteins translated from the RSVB sequenced in this study and the BA clade consensus sequence proteins. The Y axis shows positions in PRI-030 protein sequence and the X axis positions in the consensus protein sequence.

Table 4.4: Pairwise comparison of unique amino acid changes

Number of unique changes in each RSV protein. The number was calculated using multiple alignment of each PRI-024 protein with GA2 clade RSV proteins and each PRI-030 proteins with BA clade RSV proteins. An hyphen signals the absence of any unique change.

Sample	NS1	NS2	N	P	M	SH	G	F	M2-1	M2-2	L	Total unique changes
PRI-024	-	-	-	-	-	-	6	-	-	-	3	9
KF826833	-	-	-	-	-	-	1	-	-	-	1	2
KF826855	-	-	-	-	-	-	-	1	-	-	-	1
KF826849	-	-	-	-	-	-	-	-	-	-	-	0
KF530269	-	-	-	-	-	-	-	-	-	-	-	0
KF826817	-	-	-	-	1	-	1	2	-	1	-	5
KF826856	-	-	-	-	-	-	1	-	1	-	-	2
KF530264	-	1	-	-	-	-	-	-	-	1	-	2
KF530261	-	-	-	-	-	-	3	-	-	1	1	5
KF826830	-	1	1	-	1	-	2	4	1	1	1	12
KF826831	-	1	-	-	-	-	3	-	1	-	2	7
KF826838	-	-	-	-	-	-	1	-	-	-	1	2
KF530258	-	-	-	-	-	-	-	-	-	-	2	2
KF826840	-	-	-	-	-	-	1	1	-	-	2	4
KF530263	-	-	-	-	-	-	-	6	-	-	8	14
KF826848	1	1	-	-	-	-	12	5	-	5	3	27
PRI-030	-	-	-	-	-	-	2	1	-	-	4	7
KF826842	-	-	1	1	-	-	5	1	-	-	3	11
KF826822	-	1	-	-	-	-	-	-	-	-	-	1
KF826839	-	-	-	-	1	-	6	-	-	-	3	10
KF826857	-	-	-	-	-	-	-	-	-	1	1	2
KF826843	-	-	-	-	-	-	1	1	-	-	-	2
KF530262	7	-	-	-	-	-	1	1	1	-	2	12
KF826851	-	-	-	-	-	-	1	1	1	-	1	4
KF530266	-	-	-	-	-	-	-	-	-	1	1	2
KF530259	-	-	-	-	-	-	-	-	-	-	2	2
KF826845	1	1	-	-	1	-	5	1	-	-	-	9
KF826835	-	-	-	-	-	-	-	-	-	-	-	0
KF826860	-	-	-	-	-	-	3	1	-	-	1	5
KF826834	-	-	-	-	-	-	-	-	-	-	-	0
KF826844	-	-	-	-	-	-	-	-	-	-	-	0
KF826820	-	-	-	-	-	-	-	-	-	-	-	0
KF826825	-	1	1	1	-	1	2	2	1	-	5	14
KF826858	-	-	-	-	1	-	2	-	-	1	1	5
KF826829	1	1	1	-	-	-	2	2	1	2	10	20

4.4 Conclusion and discussion

Susceptibility to RSV is the result of the interplay between host, pathogen and the environment, and the pathogen genetic variability can play a role in disease outcome. Most previous RSV genomic studies focused on partial coding sequences from laboratory strains, and mainly on the attachment glycoprotein (Gpr) (*16*, *17*), which is the most variable protein in the RSV genome. Till 2011, only 6 full RSV genomes of non-laboratory strains were reported. During the past three years however, with the widespread availability of high-throughput sequencing technologies, many more RSV genomes became available. Currently there are 739 full RSV genomes available in Genbank. These data provide a valuable resource for better understanding of RSV evolution as well as development of better diagnostic test for RSV. Yet, there has been no report using full RSV sequence to study the signatures of virulence in RSV genome.

This study is a pilot exploration of the role of RSV genetic variation in the most severe clinical cases. The two RSV genomes sequenced in this study, isolated from ICU hospitalized, previously healthy children, were shown to belong to the main globally circulating clades of RSV and to be highly similar to RSV isolated from non-extreme cases at both the DNA and protein levels. We did not identify any single nucleotide or structural variant that could confer increased virulence to the virus. However, an obvious limitation of our pilot study is an extremely limited number of sequenced genomes, making it impossible to exclude an impact of RSV genomic variation in pathogenesis. We have isolated an additional 57 RSV genomes from patients with severe clinical presentation of RSV infection, for which host exome sequencing data is also available. This phenotypically homogeneous collection of RSV strains from the most extreme cases of RSV infection will provide a valuable resource to explore the role of RSV genetic variation in susceptibility and clinical outcome of the infection. Sequencing of these viruses and combining the result of viral sequencing study with the available host genomic data is needed to get a comprehensive picture of severe susceptibility to RSV. The current study provides a general framework for this research.

4.5 References

1. P. L. Collins, B. S. Graham, Viral and host factors in human respiratory syncytial virus pathogenesis., *J. Virol.* **82**, 2040–55 (2008).

2. K. J. Henrickson, S. Hoover, K. S. Kehl, W. Hua, National disease burden of respiratory viruses detected in children by polymerase chain reaction., *Pediatr. Infect. Dis. J.* **23**, S11–8 (2004).

3. L. Tan *et al.*, The comparative genomics of human respiratory syncytial virus subgroups A and B: genetic variability and molecular evolutionary dynamics., *J. Virol.* **87**, 8213–26 (2013).

4. A. Bermingham, P. L. Collins, The M2-2 protein of human respiratory syncytial virus is a regulatory factor involved in the balance between RNA replication and transcription., *Proc. Natl. Acad. Sci. U.S.A.* **96**, 11259–64 (1999).

5. M. Batonick, G. W. Wertz, Requirements for Human Respiratory Syncytial Virus Glycoproteins in Assembly and Egress from Infected Cells., *Adv Virol* **2011** (2011), doi:10.1155/2011/343408.

6. L. J. Anderson *et al.*, Antigenic characterization of respiratory syncytial virus strains with monoclonal antibodies., *J. Infect. Dis.* **151**, 626–33 (1985).

7. F. T. Bourgeois, C. Valim, A. J. McAdam, K. D. Mandl, Relative impact of influenza and respiratory syncytial virus in young children., *Pediatrics* **124**, e1072–80 (2009).

8. D. A. Rodríguez *et al.*, Predictors of severity and mortality in children hospitalized with respiratory syncytial virus infection in a tropical region., *Pediatric pulmonology* **49**, 269–76 (2014).

9. C. Sommer, B. Resch, E. Simões, Risk factors for severe respiratory syncytial virus lower respiratory tract infection., *The open microbiology journal* **5**, 144–54 (2011).

10. CB Hall, GA Weinberg, MK Iwane, The burden of respiratory syncytial virus infection in young children, *& England Journal of &* (2009), doi:10.1056/NEJMoa0804877.

11. M. Lanari, F. Prinelli, F. Adorni, S. Santo, M. Musicco, Risk factors of hospitalization for lower respiratory tract infections in infants with 33 weeks of gestational age or more: a prospective Italian cohort study on 2210 newborns, *Early Human Development* **89**, S88S90 (2013).

12. S. Parveen, S. Broor, S. K. Kapoor, K. Fowler, W. M. Sullender, Genetic diversity among respiratory syncytial viruses that have caused repeated infections in children from rural India., *J. Med. Virol.* **78**, 659–65 (2006).

13. U. J. Buchholz *et al.*, Deletion of nonstructural proteins NS1 and NS2 from pneumonia virus of mice attenuates viral replication and reduces pulmonary cytokine expression and disease., *J. Virol.* **83**, 1969–80 (2009).

14. K. L. Stokes *et al.*, Differential pathogenesis of respiratory syncytial virus clinical isolates in BALB/c mice., *J. Virol.* **85**, 5782–93 (2011).

15. M. E. Bose *et al.*, Sequencing and analysis of globally obtained human respiratory syncytial virus A and B genomes., *PLoS ONE* **10**, e0120098 (2015).

16. P. R. Johnson, M. K. Spriggs, R. A. Olmsted, P. L. Collins, The G glycoprotein of human respiratory syncytial viruses of subgroups A and B: extensive sequence divergence between antigenically related proteins, *Proceedings of the National Academy of Sciences* **84**, 5625–5629 (1987).

17. L. H. da Silva *et al.*, Genetic variability in the G protein gene of human respiratory syncytial virus isolated from the Campinas metropolitan region, Brazil., *J. Med. Virol.* **80**, 1653–60 (2008).

Chapter 5 Conclusion

Inter-individual variability is a well-known but poorly understood characteristic of infectious diseases. The work presented in this thesis provides conclusive evidence for the role of rare human genetic variants in susceptibility to life-threatening *Pseudomonas aeruginosa* (*P. aeruginosa*) and respiratory syncytial virus (RSV) infections in otherwise healthy children. In parallel, it provides a framework for exploring viral genetic variation using high-throughput sequencing (HTS) technology.

In chapter two, I tested the hypothesis that severe sepsis due to community-acquired *P. aeruginosa* in previously healthy children can reflect an underlying immunodeficiency. I used exome sequencing to search for primary immunodeficiency in 11 children with *P. aeruginosa* bacteraemia. I identified mutations leading to known immunodeficiencies in two fatal cases: X-linked agammaglobulinemia due to a novel *BTK* mutation and ICF immunodeficiency syndrome due to a novel *DNMT3B* mutation. I concluded that severe *P. aeruginosa* sepsis in children without clinical risk factors can represent the first manifestation of a previously unrecognized primary immunodeficiency.

In chapter three, I tested the hypothesis that rare genetic variations can explain the clinical severity of common viral respiratory infections in previously healthy children. 120 previously healthy children requiring intensive care support because of a severe illness caused by a respiratory virus were recruited. After exome sequencing, I identified three rare loss-of-function variants in *IFIH1*. Functional characterization of the variants demonstrated that the mutated proteins are unable to induce interferon-β, are intrinsically less stable than wild-type IFIH1, and lack ATPase activity. Experiments also showed a dominant negative effect of the mutated proteins on wild-type IFIH1 stability and ATPase activity. *In vitro* assays showed effective restriction of human respiratory syncytial virus and rhinovirus by IFIH1. I concluded that *IFIH1* deficiency causes a primary immunodeficiency manifested in extreme susceptibility to common respiratory RNA viruses.

In chapter four, I performed a pilot study for exploring RSV genetic variation. The main goal of this work was to set up a framework for analysing HTS data derived from the pathogen genome. Two RSV strains were isolated from nasopharyngeal aspirates (NPA) of two previously healthy children with severe clinical presentation of RSV infection. After sequencing, I used bioinformatic analysis to reconstruct the RSV genome for each strain and to compare them at the RNA and protein level with strains isolated from routine cases of RSV infection. The analysis showed that these two RSV genomes belonged to the most prevalent clades. I did not find any variation at RNA or protein level that can explain the unusual severe outcome of RSV infection in the selected patients. However, given the high variability of RSV genome and the small size of this study, one cannot exclude an impact of RSV genomic variation on pathogenesis and clinical outcome.

The results presented in chapters two and three of this thesis demonstrate that exome-sequencing is a powerful tool to study the genetic basis of infectious disorders and highlights the power of extreme phenotype sampling in sporadic cases for gene discovery. The prototype study I described in chapter four can be extended to large-scale studies aiming at establishing a link between pathogen genome variation and disease severity. Pathogen genomic studies in combination with host genomic studies can provide a more comprehensive picture of host-pathogen interaction and consequently lead to development of better preventive and therapeutic measures.

In our host genomics studies, causal mutations were found and confirmed in 18% (for *P. aeruginosa* infection) and 7% (for RSV infection) of cases. While non-genetic and pathogen-related factors could explain some of the remaining cases, we cannot rule out the possibility of other human genetic factors being involved. Larger sample sizes in combination with functional follow-up of new candidate variants is required to find pathogenic variants in these cases. Furthermore, whole genome sequencing is needed to explore the non-coding variants, large structural variants and also exonic variants that might not be well covered using current exome capture kits. An obvious limitation of our pathogen genomics study is the extremely limited number of sequenced genomes. We have isolated an additional 57 RSV genomes from patients with severe clinical presentation of RSV infection. Sequencing of these additional RSV genomes and combining the results with our data from host genomics of RSV infection is a necessary next step in our efforts to study the role of RSV genetic diversity in disease severity.

Although the role of heritable factors in infectious diseases has been suspected for a long time, it’s only during the past decade that multiple studies were published that demonstrated the importance of host genetics in the pathogenesis of several infectious diseases. During the same period, the maturation and democratization of high-throughput sequencing technologies has made it possible to run pathogen genomic studies that have widened and sharpened our understanding of the genetic diversity of pathogens, a necessary step toward deciphering the role of pathogen genetic variability in disease pathogenesis. Because of the unremitting nature of the evolutionary arms race between host and pathogen, infectious diseases will continue to be part of the human experience. By promising a more fundamental understanding of the genetic contributors to infection outcomes, the current genomic era has the potential to bring better preventative and therapeutic strategies to individuals and populations.

Curriculum Vitae

Samira Asgari

PhD student
EPFL station 19, SV-GHI-GRFE
1015 Lausanne, Switzerland
samira.asgari@epfl.ch
+41 (0) 21 69 31 874

Education

2011-2016:	PhD program in biotechnology and bioengineering, École Polytechnique Fédéral de Lausanne (EPFL), Switzerland
2004-2011:	MS in Medical Biotechnology, University of Tehran, Iran

Research experience

2011-present: PhD project, EPFL, Switzerland

Host genomics of infectious disease

- Susceptibility to viral respiratory infections in children
- Septic shock due to Pseudomonas aeruginosa in children
- Fulminant hepatitis B in adults

Pathogen genomics

- Deep sequencing and analysis of human Respiratory Syncytial virus (hRSV) genome

2009-2011: MS project, Royan Institute, Iran

Directed differentiation of human induced pluripotent stem cells into functional hepatocytes and their transplantation to mouse model of acute liver failure

Skills

Computational

Analysis of next generation sequencing data (exome and RNA sequencing), Genomics software (plink, BWA, Bowtie, SAMtools, PICARD, GATK, SnpEff, SIFT, PolyPhen, SKAT, VEP, ANOVAR, etc.), Biological database (1000GP, HapMap, UCSC, EVS, NCBI, ExAC, UK10K, etc.), Adobe illustrator, Linux OS, shell and C-shell scripting, R, familiar with python

Experimental

Cell culture, cell viability assays, flow cytometry, gel electrophoresis, cloning recombinant proteins; western blotting; ELISA, affinity chromatography, familiar with many biochemical and biophysical chemistry techniques

Soft skills

Teaching and project supervision

2014:	Summer project supervision, EPFL, Fellay lab
2014:	Master project supervision, EPFL, Fellay lab

2012-2013: 175 hours of teaching assistantship, EPFL, School of Life Sciences

Communication and Management

2014: Executive member of the "BioScience Network Lausanne"
2013: President of "EPFL/UNIL Toastmasters club"
2013: Competent communicator, awarded by Toastmasters international

Language

Persian: Mother tongue, English: C2, French: C1

Awards

2015: ASHG/Charles J. Epstein Trainee Award for Excellence in Human Genetics Research – **Finalist**, Baltimore MD, USA
2015: International Primary Immunodeficiencies Congress, **Best abstract award**, Budapest, Hungary
2014: EPFL Nominee for Global Young Scientists Summit@one-north 2014, Singapore
2009-2011: National Elite Students Foundation of Iran fellowship
2004-2009: Research and Technology fellowship for Exceptional Talents, Ministry of Science, Research and Technology, Iran

Conferences

2015: Oral presentation, American Society of Human Genetics annual meeting, Baltimore, MD
2015: Oral presentation, International Primary Immunodeficiencies Congress, Budapest, Hungary
2015: Oral presentation, Childhood Life-threatening Infectious Disease meeting, Siena, Italy
2015: Oral presentation, Basel Conference in Computational Biology, Basel, Switzerland
2014: Oral presentation, Leena Peltonen School of Human Genomics, Welcome Trust Genome Campus, Hinxton, Cambridge, UK
2012: Oral presentation, Young Investigator's Forum, European Association for Study of Liver annual meeting, Barcelona, Spain

Publications

- **Asgari, S.**, Schlapbach, LJ., Anchisi, S., Hammer, C., Bartha, I., and Fellay, J. Loss-of-function mutations in *IFIH1* predispose to severe viral respiratory infections in children. **Asgari, S.**, Schlapbach, LJ., McLaren, PJ. Bartha, I., Wong, M., and Fellay, J. Exome sequencing reveals primary immunodeficiencies in children with fulminant community-acquired pseudomonas aeruginosa sepsis. Submitted
- **Asgari, S.**, Moslem, M., Bagheri-Lankarani, K., Pournasr, B., Miryounesi, M., and Baharvand, H. (2013). Differentiation and transplantation of human induced pluripotent stem cell-derived hepatocyte-like cells. Stem Cell Rev *9*, 493–504.
- **Asgari, S.**, Pournasr, B., Salekdeh, G.H., Ghodsizadeh, A., Ott, M., and Baharvand, H.

(2010). Induced pluripotent stem cells: a new era for hepatology. J. Hepatol. *53*, 738–751.

- Fattahi, F., **Asgari, S.**, Pournasr, B., Seifinejad, A., Totonchi, M., Taei, A., Aghdami, N., Salekdeh, G.H., and Baharvand, H. (2013). Disease-corrected hepatocyte-like cells from familial hypercholesterolemia-induced pluripotent stem cells. Mol. Biotechnol. *54*, 863–873.
- Translation: Sadeghi, M., **Asgari, S.**, Falahati, F., Ghodsizadeh, A., Jadaliha, M., Post-genomic informatics by Kanehisa Minoru, Tehran, National Institute for Genetic Engineering and Biotechnology, 2011