

University of Groningen

## Mapping complex and monogenetic disorders

Szperl, Agata Maria

**IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.**

*Document Version*

Publisher's PDF, also known as Version of record

*Publication date:*

2012

[Link to publication in University of Groningen/UMCG research database](#)

*Citation for published version (APA):*

Szperl, A. M. (2012). *Mapping complex and monogenetic disorders: methods and applications*. s.n.

### Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

### Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

RIJKSUNIVERSITEIT GRONINGEN

**Mapping complex and monogenetic disorders:  
methods and applications**

Proefschrift

ter verkrijging van het doctoraat in de  
Medische Wetenschappen  
aan de Rijksuniversiteit Groningen  
op gezag van de  
Rector Magnificus, dr. E. Sterken,  
in het openbaar te verdedigen op  
woensdag 12 september 2012  
om 12.45 uur

door

**Agata Maria Szperl**

geboren op 16 juni 1982  
te Kielce, Polen

Promotor:

Prof. dr. C. Wijmenga

Beoordelingscommissie:

Prof. dr. A.M.H. Boots

Prof. dr. P.C. Limburg

Prof. dr. A.J. Moshage

Prof. dr. E.H.H.M. Rings

ISBN: 978-90-367-5721-8

Design and printing:



Lovebird design & printing solutions  
[www.lovebird-design.com](http://www.lovebird-design.com)



## Table of contents

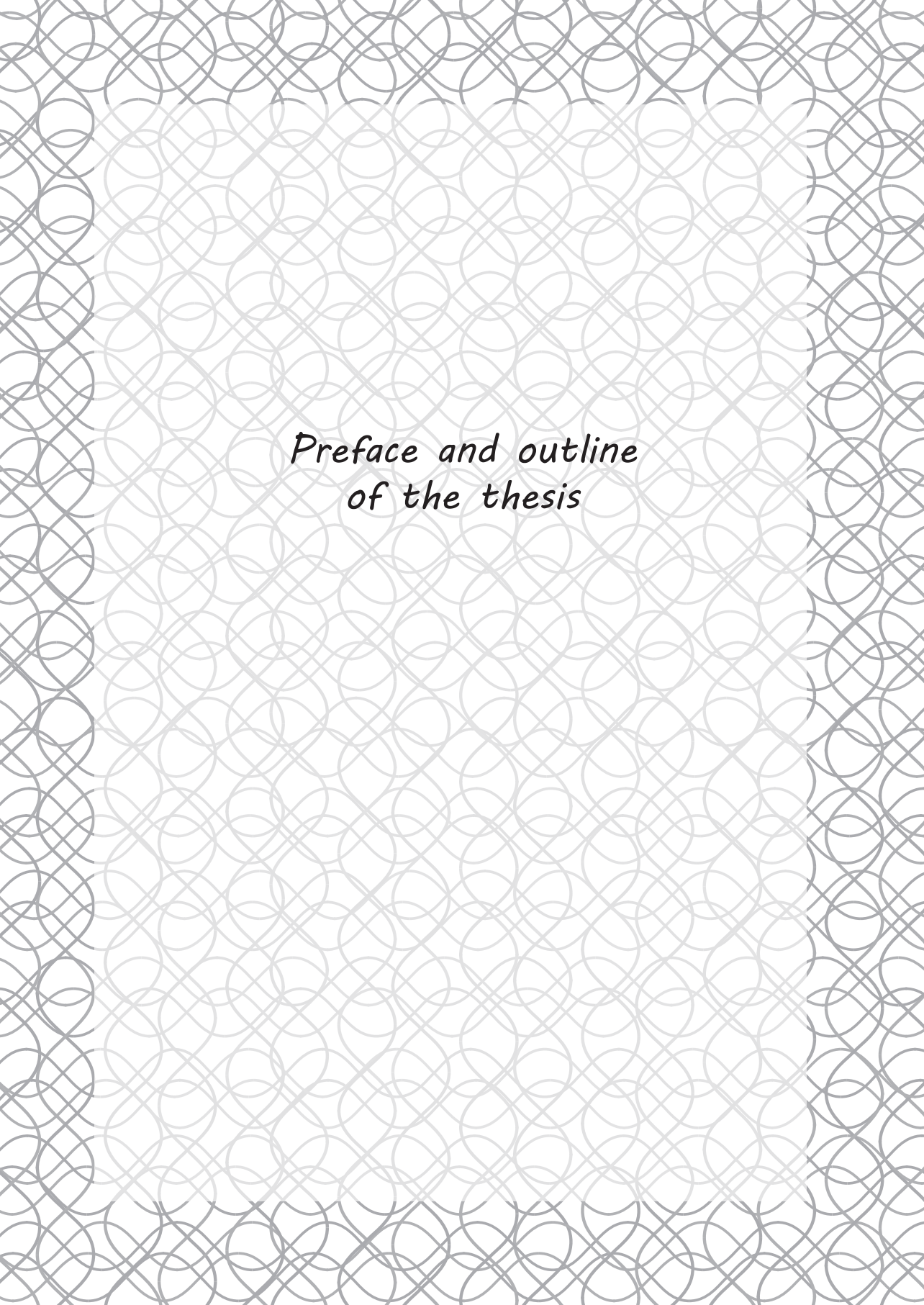
*Preface and outline of the thesis*

*Abbreviations*

<i>Chapter I</i>	Introduction: Mendelian vs. complex disorders
<i>Chapter II</i>	Functional characterization of mutations in the myosin Vb gene associated with microvillus inclusion disease. <i>Journal of Pediatric Gastroenterology and Nutrition</i> , 2011, 52(3):307-13
<i>Chapter III</i>	True autosomal dominant inheritance of FMF caused by a mutation in exon 8 of the MEFV gene. In preparation
<i>Chapter IV</i>	Exome sequencing in a family segregating for celiac disease. <i>Clinical Genetics</i> , 2011, 80: 138-147
<i>Chapter V</i>	Functional polymorphism in IL12B promoter site is associated with ulcerative colitis. <i>Inflammatory Bowel Disease</i> , 2011, 17(6):E38-40
<i>Chapter VI</i>	Cross-ethnic replication and fine-mapping of coeliac disease loci in north Indian population. In preparation
<i>Chapter VII</i>	Discussion and future perspectives
<i>Summary in Dutch</i>	
<i>Summary in English</i>	
<i>Summary in Polish</i>	
<i>Acknowledgments</i>	
<i>Curriculum Vitae</i>	



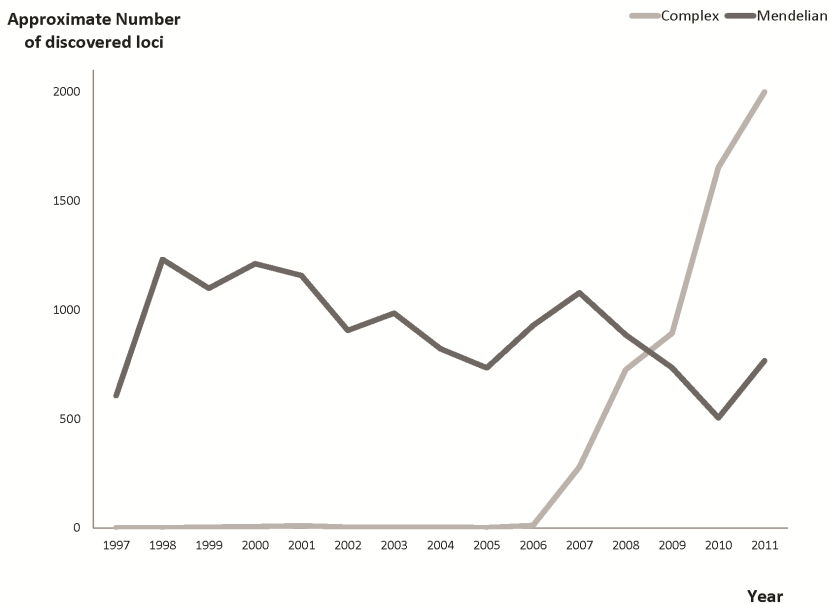




*Preface and outline  
of the thesis*



The majority of diseases are caused or influenced by hereditary factors. Most of the rare disorders, which are often devastating and sometimes even life-threatening, are monogenetic disorders caused by rare genetic changes (mutations). In contrast, complex diseases, which are often chronic diseases or diseases of the elderly, are caused by the interplay between rather common genetic variants and environmental factors. To understand, predict and possibly cure hereditary diseases, recognition of the causal genetic factors is crucial. The main strategy that is followed in the discovery of causative variants for both rare and common diseases, is to first identify candidate loci (i.e. a region in genome either associated or linked to the disease under study and containing one or more genes), then fine-mapping of the region (to refine the region to preferably a single gene), followed by investigation of the refined candidate locus for causative mutations by direct re-sequencing. When we look at the progress made in disease gene mapping from 1997 onwards (Figure 1), we can see a fairly steady and high discovery rate in the field of rare disorders.



**Figure 1** Approximate number of mapped loci from 1997 till 2011 for Mendelian and complex diseases.

However, success in mapping complex traits was lagging behind, mainly due to the methods available. Linkage studies, for example, have proved extremely powerful for the mapping of monogenetic disorders, but have hardly contributed to the discovery of common loci. The field of complex disease genetics changed entirely in 2005 with the introduction of the Common Disease-Common Variant hypothesis and



the availability of new tools to test it. At that time the Human Genome Project had delivered a Human Catalogue of Common Variants and inspired the development of array technology to include these variants on a chip for large-scale genome-wide genotyping. This laid the foundation for association studies in complex traits. Since the first genome-wide association study (GWAS) in 2006, thousands of loci have been correlated with hundreds of complex traits.

For many years fine-mapping of the candidate loci was an absolute necessity before embarking on the re-sequencing of single genes due to the lack of affordable and high-throughput methods. Again, it was a technological breakthrough that changed the field of gene mapping. In 2009 next generation sequencing technology was launched, a revolutionary method whereby resequencing of the entire human genome could be done within a week compared to the 13 years it took to complete the Human Genome Project. At the same time, the cost had also dropped from \$3 billion to approximately \$15,000 per genome. Next generation sequencing quickly became an attractive tool for investigating candidate regions, as well as for sequencing all the genes in the genome and solving a disease without previous mapping information (i.e. exome sequencing). To date, many rare, monogenetic disorders (recessive and dominant), as well as a limited number of complex diseases, have been resolved using this approach.

The aim of this thesis was to use different state-of-the art approaches for fine-mapping and discovering new genetic variants causing both monogenetic and complex diseases.

In **chapter 1**, I briefly present the history of genetic mapping and explain the differences between rare and complex diseases with respect to the genetic and environmental factors as well as the nature of the variants. I further describe methods used for mapping candidate regions in complex and rare disorders and factors that could influence the replication phase. Finally I explain the powerful tool of next generation sequencing and its possible applications in discovering new variants and fine-mapping candidate regions.

In **chapter 2**, “Functional characterization of mutations in the myosin Vb gene associated with microvillus inclusion disease”, I present the wide spectrum of mutations in the myosin Vb gene that lead to microvillus inclusion disease (MVID). In this study I show three families of different ethnic origins in which the same disease is caused by different mutations in the same gene (i.e. allelic heterogeneity). I also

demonstrate the power of combining homozygosity mapping with the knowledge of the protein product in uncovering the candidate gene.

In **chapter 3**, “True autosomal dominant inheritance of FMF caused by a mutation in exon 8 of the *MEFV* gene”, I present an exome-wide sequencing study as a mutation mapping tool for this dominant autosomal, periodic fever syndrome. By sequencing two affected, related individuals and applying a “step-wise” filtering approach, I was able to map the causative mutation to the 577 codon in the *MVEF* gene. This work clearly shows the power and high sensitivity of next generation sequencing in looking for causative mutations in monogenic diseases.

**Chapter 4**, “Exome sequencing in a family segregating for celiac disease”, presents the use of next generation sequencing for fine mapping linkage regions in a four-generation family segregating for celiac disease. This complex disease presented in a dominant-like manner. As we were not able to find a likely candidate variant in any of the linkage regions, we also investigated the entire exome for nonsense variants that could explain the heritability of the disease in this family. Although we were not able to identify a mutation, we pinpointed the different aspects and bottlenecks that show how difficult it is to solve a complex disease in familial studies.

In **chapter 5**, “Functional polymorphism in IL12B promoter site is associated with ulcerative colitis”, I show the correlation of genetic variants in the *IL12b* gene to ulcerative colitis, one of the subtypes of Inflammatory Bowel Disease, which is a complex immune-mediated disease. This study presents a way of mapping causative variants by the “cross-disease” replication of a variant in a gene that has already been associated to another immune-mediated disease.

In **chapter 6**, “Cross-ethnic replication and fine-mapping of celiac disease loci in a north Indian population”, we performed a cross-ethnic replication study in a north Indian population of 26 loci previously associated to celiac disease in Europeans. To correct for the high allelic heterogeneity caused by differences in linkage disequilibrium structure between Europeans and Indians, we used two replication methods and successfully replicated 50% of the association. Furthermore, we were able to narrow down the association signal using an innovative Cross test statistic to check for risk haplotypes at the *IL2-IL21* locus, since the same risk haplotype in Indians was shorter due to the smaller linkage disequilibrium intervals in that population.

In **chapter 7**, I summarize all the work I have done and discuss the future perspectives for the mapping of causative variants in monogenic and complex diseases.

BP	Base-pair
CD-CV	Common disease-common variant
CD-RV	Common disease-rare variant
CeD	Celiac disease
cM	Centimorgan
CNV	Copy number variation
dbSNP	The Single Nucleotide Polymorphism Database
DNA	Deoxyribonucleic acid
ESP	Exome sequencing project
FMF	Familial Mediterranean fever
GIH	Gujarati Indians in Houston
GoNL	Genome of the Netherlands project
GWAS	Genome-wide association study
HapMap	Haplotype map of human genome
HLA	Human leucocyte antigen
IBD	Inflammatory bowel disease
ICF	Immunoglobulin deficiency syndrome
IL	Interleukine
Indels	Insertion/deletion
Kb	Kilobases
LD	Linkage disequilibrium
LOD	Logarithm of the odds
MAF	Minor allele frequency
Mb	Megabase
MDS	Multi-dimensional scaling analysis

MEFV	Mediterranean fever gene
MVID	Microvillus inclusion disease
MYO5b	Myosin V b gene
NGS	Next generation sequencing
NPL	Non-parametrical linkage values
OMIM	Online Mendelian Inheritance in Man
OR	Odds ratio
PBMC	Peripheral-blood mononuclear cells
PCR	Polymerase chain reaction
qRT-PCR	Quantative real-time polymerase chain reaction
RNA	Ribonucleic acid
SNP	Single nucleotide polymorphism
SV	Sequence variant
UC	Ulcerative colitis
UTR	Untranslated region



The background of the slide is a repeating pattern of overlapping circles, creating a mesh-like texture. The circles are light gray and overlap in a way that creates a series of smaller, interconnected shapes.

## *Chapter 1*

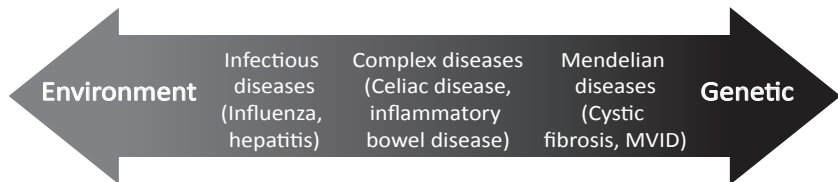
*Introduction:  
Monogenic vs complex  
disorders*



## Mendelian vs. complex disorders

There are about 25,000 protein-coding genes in the human genome and any alteration in these genes can potentially lead to a genetic disorder. The main concept underlying a genetic disorder is that of inheritance; the disease may affect other people related to the **proband** and it will **co-segregate** in families. One of the first genetic diseases to be recognized was reported in the 2<sup>nd</sup> century by Rabbi Simon and Gamaliel. They described a family in which an infant and its three maternal cousins had died from extensive bleeding after medical procedures; this disorder is now known as hemophilia and it is caused by **mutations** (alterations) in genes regulating the blood coagulation (1).

Some genetic disorders are strongly determined by genes, these are called monogenic or Mendelian disorders, while many other disorders result from multiple genes interacting with environmental factors (called multifactorial or complex disorders). Thus, we can place each disease at a different point along a continuous spectrum, depending on whether the genetic or environmental effect is the strongest determining factor. The majority of infectious diseases are triggered by environmental factors so these lie at one extreme of the spectrum (Figure 1). Mendelian diseases, such as cystic fibrosis, lie at the other extreme and are mainly determined by mutations in the genome. Complex disorders lie in the middle of the spectrum as they are strongly influenced by both types of factors (Figure 1).



**Figure 1 The continuous spectrum of disease causation.** Mendelian diseases, like cystic fibrosis or microvillus inclusion disease (MVID), are caused by strong genetic factors, whereas other disorders are similar to infectious diseases and are mainly caused by environmental factors. Complex disorders lie in the middle of the spectrum and are due to environmental and genetic factors interacting and playing an important role in the pathogenesis.

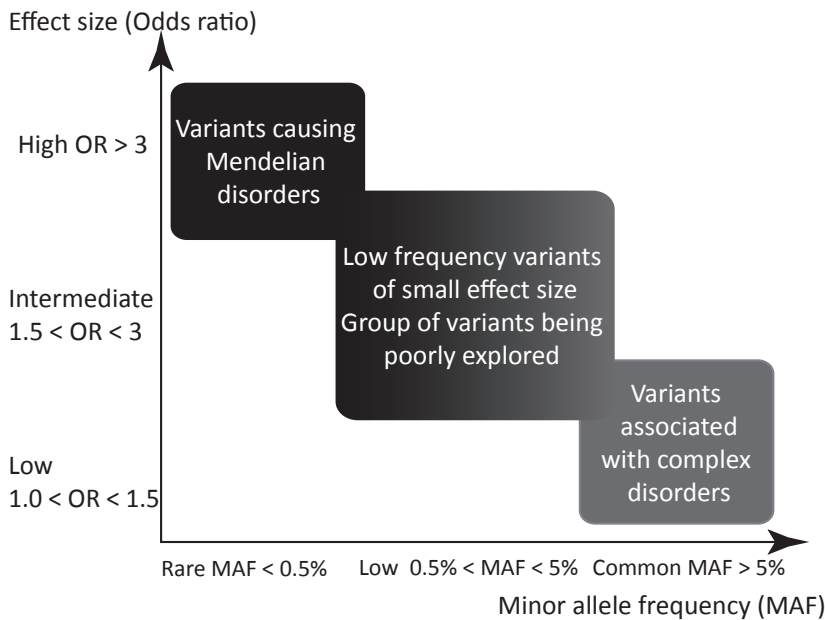
**Proband:** is the first affected family member who seeks medical attention and was diagnosed with a genetic disorder.

**Co-segregation:** is a segregation pattern that shows how often the variant coincides with the disorder in the pedigree.

**Mutation:** is an alteration in a gene from its natural state.



Monogenic disorders are rare, and caused by mutations of a large **effect size** in one gene (Figure 2). These mutations result in very low **minor allele frequencies (MAF)** of the mutant alleles ( $MAF < 1\%$ ). The mutations often occur in highly **conserved regions** that are prone to negative evolutionary selection (2, 3). Around 60% of the genetic mutations leading to monogenic disorders have a **missense or nonsense** character, around 10% are **splice sites**, 7% are insertions/deletions (**indels**) and less than 1% occur in regulatory regions (4).



**Figure 2** There is a wide range of variants with different effect sizes (**OR, odds ratio**) and minor allele frequencies (**MAF**). Mendelian diseases are caused by rare variants with very low MAF and a high OR; they can be successfully identified by linkage or homozygosity mapping, for example. Common variants with high MAF and a low OR are often associated to complex diseases. Low frequency variants of intermediate effect size are an unexplored group as they are not detectable with **linkage** or **genome-wide association studies (GWAS)**. One way to investigate their involvement in genetic disease is to sequence the entire genome, which has become feasible with next-generation sequencing technology.

**Effect size:** the strength of relation between variant and the trait.

**Minor-allele frequency (MAF):** the frequency of the allele that is less often observed in a general population.

**Conserved regions:** are crucial and functionally important regions in DNA that do not allow variation by strong negative selection.

**Missense mutation:** is a change in the base sequence of a gene that alters or eliminates a protein.

**Nonsense mutation:** is when codon is changed to a stop codon, resulting in a truncated protein product.

**Splice site mutation:** is when alternation is location in the DNA sequence involved in removing the noncoding areas to form a continuous gene transcript for translation into a protein (so called splicing).

**Indels:** are defined as either an insertion or deletion. Insertion involves the addition of genetic material whereas deletion lack of the genetic material.

**Odds ratio (OR):** is a statistical measure of effect size reflecting the effect of the mutation on the disease

**Linkage or homozygosity mapping:** are family-based methods used to find the region in a genome carrying the causative gene. Widely used for mapping monogenic disorders.

**Genome-wide association study (GWAS):** is population-based method (case-control) for mapping regions associated to a disease.

Widely used for mapping complex traits.

Gregor Mendel was, in 1866, one of the first to outline the inheritance of monogenic traits in pea plants, and he described recessive and dominant models. In 1902, Garrod showed that alkaptonuria (black urine disease) was inherited recessively, so that the “inborn errors of metabolism” were present in the proband’s siblings but not in the parents (5). The inheritance of monogenic disorders has been well established and follows six main patterns:

**Dominant:** alleles that determine the phenotype displayed in a heterozygote with another (recessive) allele.

**Recessive:** is a gene that is phenotypically manifest in the homozygous state but is masked in the presence of a dominant allele.

**Allele:** is an alternative form of a gene; any one of several mutational forms of a gene.

**Autosome:** is a nuclear chromosome other than the X- and Y-chromosomes.

**Mitochondrial DNA:** is a mitochondrial genome consists of a circular DNA duplex, with 5 to 10 copies per organelle.

**Celiac disease (CeD):** is an complex disease with very strong immune response, in small intestine, to the gluten proteins.

**Polymorphism:** is a change in the genomic DNA and it differs from a reference sequence.

- 1) **Autosomal dominant:** where the presence of one mutated **allele** on an **autosome** is enough to cause the disease,
- 2) **Autosomal recessive:** where the presence of two mutated alleles in the same gene on two homologous autosomes leads to disease,
- 3) **X-linked recessive:** where the presence of two mutated alleles on the X chromosome (in females) or one (in males) are needed to cause disease,
- 4) **X-linked dominant:** where the presence of one mutated allele on the X chromosome is enough to cause disease,
- 5) **Y-linked:** where the presence of one mutated allele on the Y chromosome is needed to cause disease,
- 6) **Mitochondrial-linked:** where maternal **mitochondrial DNA** carries the mutation causing the disease. This is not typical Mendelian inheritance, as none of the parental DNA is inherited, however, these disorders are also monogenic.

Complex diseases are fairly common in the general population. For example, the prevalence of **celiac disease**, a complex, immune-related intolerance to gluten, is around 1% in Caucasian populations (6). The first hypothesis on the character of causative variants for complex diseases was based on their high prevalence in the population, and was called the “common disease-common variant” (CD-CV) hypothesis (Figure 3). It states that complex diseases are caused by a combination of multiple common **polymorphisms**, with MAF > 5%, of small effect

size, with each polymorphism explaining a small part of the heritability (Figure 2) (7, 8). In total, from 2005 to 2010, some 5,854 common genetic **single nucleotide polymorphisms (SNPs)** were found to be associated with 540 complex disorders (9). Celiac disease is currently associated with 57 non-HLA SNPs representing 39 loci, and although no causative gene was found, the majority of these SNPs map close to immune-related genes, explaining some 15% of the **heritability** (10). The work of Thomas and Kejariwal showed that the majority of **coding variants** associated through GWAS studies with complex diseases map to less conserved regions than the severe mutations found for Mendelian diseases, which may explain their more moderate effect size (3).

A second hypothesis, “common disease-rare variant” (CD-RV), states that complex genetic diseases could be due to variants of low MAF (< 5%), and moderate effect size (8) (Figure 2), with an incomplete **penetrance** in some families (11). The CD-RV hypothesis could explain part of the “hidden heritability” of complex disorders. Moderate and rare variants have so far remained largely undiscovered, as the DNA chips used for genome-wide association studies (GWAS) do not contain variants with a MAF < 5% (8, 12). These variants could be present within the GWAS loci and collectively drive the association to the disease (“**synthetic association**”) (13, 14) (Figure 3). For example, several rare variants in the *NOD2* gene are strongly associated with **inflammatory bowel disease (IBD)** and although the variants show a familial clustering they do not follow a clear inheritance pattern (11). The work of Cohen et al. in 2004 and 2005 showed that extreme phenotypes of metabolic traits are also strongly correlated with rare variants in metabolic candidate genes (15, 16). The CD-RV hypothesis is very likely to hold true for the majority of familial, Mendelian-like subtypes of complex disorders. For example, a mutation in the *VPS35* gene has been identified in a large Austrian family segregating for Familial Parkinson’s disease (17), and recent **case-control studies** on hyperglycemia, IBD and gout have shown an excess of rare variants in loci mapped by GWAS (18-20).

**Single Nucleotide Polymorphism (SNP):** is a DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence is altered.

**Heritability:** the proportion of the phenotypic variation in a trait that is due only to the genetic set up. It is measure as the correlation of traits between related individuals. The part of the heritability that cannot be explained with genome-wide association studies for complex diseases is called hidden heritability.

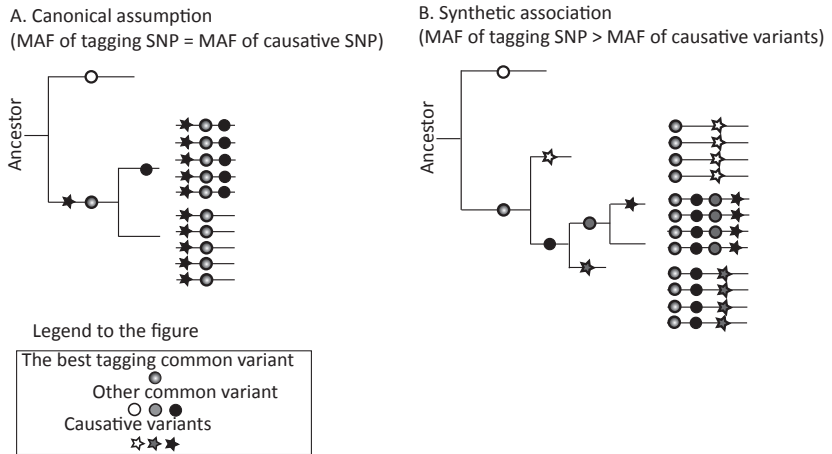
**Coding variants:** are changes in the part of the genome coding for proteins.

**Penetrance:** is the probability of the genetic trait to be expressed.

**Synthetic association hypothesis:** is one of the explanations for hidden heritability of complex diseases. It claims that cluster of low-frequency variants drive the association of common loci in GWA studies.

**Inflammatory bowel disease (IBD):** is an autoimmune, complex disorder affecting either the small intestine (ulcerative colitis) or the entire gastrointestinal tract (Crohn’s disease).

**Case-control design:** is a population study where the hypothesis is tested by comparing occurrence of the object (e.g. frequency of an allele) between cases (diseased subjects) and controls (non-diseased subjects).



**Figure 3** Illustration presents genealogical tree showing rising of the tagging (○●●●) and causative variants (☆☆☆). Panel A represents common disease-common variant (CD-CV) theory where the best tagging variant (○) captures the causative common variant (●). Panel B represents “synthetic association” hypothesis where the best tagging SNP (☆) captures the combined effect of many causal rare variants (☆☆☆). Figure adapted with permission from Wang et al., 2010 (39).

In summary, the current data suggests that both hypotheses, CD-CV and CD-RV, may hold true and help explain the genetic architecture of complex diseases, since there is a wide genetic heterogeneity within the variants so far associated to complex diseases with regard to their effect size and frequency. The fractions of common and rare variants explaining the heritability of complex disorders will also vary for each disease (12, 21).

### Hunting for the causative genes

To understand and treat genetic disorders, it is essential to identify the causal gene or variant. In general, there are two ways of looking for the causal gene: using a hypothesis-driven (a) or a hypothesis-free approach (b). In a genome-wide approach, we do not apply any hypothesis for searching for certain candidate genes or loci (22). In a hypothesis-driven approach, on the other hand, the candidate gene is investigated based on known, underlying disease mechanisms, independent of the gene mapping strategy used (23) (Figure 4).

### (a) Hypothesis-driven approaches

In hypothesis-driven approaches, knowledge of underlying mechanisms of the disease suggests protein products for possible involvement in the disease pathogenesis (“reverse genetics”). For example, Sato et al. (2007) showed that Rab-8-deficient mice expressed a reduced amount of the RAB8 protein, identical to what was observed in a patient with microvillus inclusion disease (MVID) (24). Although *Rab8* turned out not to be the causative gene for MVID, this experiment still suggested further candidate genes for screening and led to the discovery of *MYO5b* as the causal gene that interacts with RAB8 in the same pathway (25, 26). Using reverse genetics without mapping the candidate **locus** has severe limitations, especially in identifying genes for complex disorders, since these can have a broad phenotypic spectrum and thus an unclear disease mechanism. A candidate approach is therefore most effective when applied to a previously mapped locus (so-called “positional candidate mapping”).

**Locus:** is the physical place in the DNA sequence measured in base-pairs (bp).

### (b) Hypothesis-free approach

The hypothesis-free approach can be divided into two phases: first a genome-wide scan, in which appropriate statistical measures are applied to determine any significant association between the gene’s position in the genome and the disease (by either linkage analysis, homozygosity mapping, or a genome-wide association study), or by sequencing the entire genome for variants matching the disease criteria (whole genome re-sequencing). The second phase is replication, where the region, locus or variant is investigated in an independent study designed to confirm the preliminary findings.

#### Genome-wide mapping phase

Genome-wide linkage scans These are mainly used to map candidate loci in families segregating for disease, by finding significant linkage between a marker (for example, a SNP or **microsatellite**) and the disease in the pedigree (family-based), while taking **sequence variants** (SVs) into account. There are two types of linkage analysis:

**Microsatellite:** also called short tandem repeats, are highly polymorphic sequences used for the genetic mapping.

**Sequence variant (SV):** is a change in the genomic DNA found by re-sequencing methods that differs from a reference sequence.

first, model-based linkage (parametrical) in which a Mendelian model of inheritance is assumed for analyzing the co-segregation. This model can be recessive or dominant. In principle, co-segregation decreases if two loci are far away from each other, so that **recombination** takes place more frequently (27). The significance of parametrical linkage is reported as a logarithm of the odds (LOD) score; this is a function of recombination for each genotyped locus (or loci) at the disease locus. The larger the LOD score, the stronger the evidence for linkage with the disease, but a negative LOD score means there is no co-segregation present. This approach was successfully used in mapping Mendelian candidate loci with a strong genotype-phenotype correlation.

The second type of linkage analysis is non-parametrical linkage in which no model assumptions are made. This type of linkage is used in complex diseases, since there is often an incomplete penetrance. The analysis is performed in only the affected individuals from the family (“affected only”); the linkage is based on the sharing of alleles, **identical-by-descent**, at the disease locus. Both parametrical and model-free approaches have been used in mapping the candidate regions in large families that segregate for a complex disease (28, 29). Before the GWAS era (i.e. pre-2005), the most widely used method for mapping complex disorders was a linkage-based analysis in **sib-pairs**, where both first-degree relatives are affected (27, 30). There is 0.25 probability of no identity-by-descent sharing, 0.5 probability for one allele to be shared, and 0.25 for two alleles to be shared as identity-by-descent. A linkage approach would be adopted if both sibs share more identity-by-descent than would be expected by chance. Such an approach was used to map a candidate locus on chromosome 19 linked to celiac disease (31).

Genome-wide homozygosity mapping This approach is used to study rare, autosomal recessive diseases by finding large, homozygous regions spanning the causative mutation in **consanguineous families** and in inbred populations (32). In recessive diseases, the children of related parents (for example, parents who are first cousins), often carry the causative mutation within large, homozygous-by-descent loci

**Recombination:** is the exchange of DNA fragments during meiosis.

**Identical-by-descent:** is when the inheritance of allele at locus can be established vertically whereas identical-by-state (IBS) is when the alleles are the same but the inheritance is unknown.

**Sib-pairs:** are brother-brother, sister-brother or sister-sister.

**Consanguineous families:** where the parents in a family are blood relations (inbreeding) often first cousins. Populations with a high degree of consanguinity are called inbred populations.

(autozygosity). In complete pedigrees, we can apply the same statistical techniques as for parametrical linkage and the LOD score will indicate whether the linkage is significant (32). An alternative approach is to genotype affected siblings from the family and to look for the largest homozygosity region, as it has been shown that causative mutations often map to one of the larger homozygous regions (26, 33).

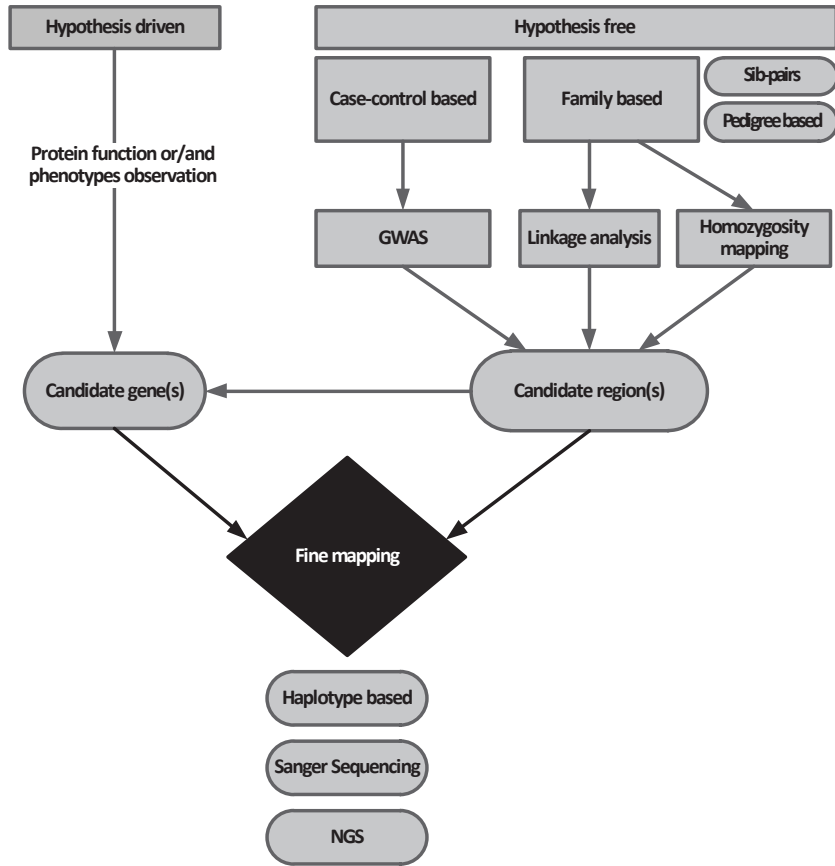
Genome-wide association studies (GWAS) These studies are used to investigate complex traits by finding associations between SNPs and the disease, mostly by using a case-control design. In GWAS, the significance of the differences in allele frequencies of genotyped markers is established in controls (unaffected individuals) versus cases (affected individuals) (34, 35). Markers used for association studies are selected from common SNPs in **HapMap**, which is an international catalogue of genetic variants in humans (36). Hence, GWAS can be used to test the CV-CD hypothesis for any complex disease (22). Due to the small effect size of the individual risk variants ( $OR < 1.5$ ), a high number of associated loci explain only a small part of the heritability (see Figure 2). For example, around 71 loci have been associated with Crohn's disease (a subtype of inflammatory bowel disease), but explain only 20% of the heritable risk (37). A standard GWAS is an indirect mapping approach where an associated, common SNP ( $MAF > 5\%$ ) is tagging (**tagging SNP**) the unknown, causal variant. This is possible since variants are in **linkage disequilibrium (LD)**, which means they are inter-dependent, create **haplotypes**, and are inherited together. Thus, a genotyped SNP can predict and 'tag' other variants that have not been genotyped (34, 38). As a GWAS can identify loci rather than single genes, further fine-mapping of the associated loci is required to find the true causal gene or variant (39). There are two hypotheses on the character of causative, tagged variants: one is that the common SNP is tagging the other common causative variant of small effect (CD-CV hypothesis), when a causal mutation arose together with the common tagging SNP; the other is where the common SNP tags rarer variants that arose after the tagging SNP became common (synthetic association) (Figure 2). Both these hypotheses were discussed earlier.

The International **HapMap Project**: is an international effort to identify and catalog genetic similarities and differences in human beings. It describes what these variants are, where they occur in our DNA, and how they are distributed among people within populations and among populations.

**Tagging SNP**: is a common polymorphism that is in high LD with other variants and represents them. These SNPs are used in GWAS to map regions of association to disease.

**Linkage disequilibrium (LD)**: is the association between two loci and reflects the probability for variants being inherited together more often than would be expected by chance.

**Haplotype**: the sequence of alleles inherited together.



**Figure 4** Scheme of steps taken to identify candidate variants. Hypothesis-driven: knowledge of the disease is used to find the candidate gene(s). Hypothesis-free: the candidate region is mapped. Information on the phenotype or protein function can be applied in the positional candidate approach. The candidate genes or regions (from hypothesis-free and hypothesis-driven approaches) are screened for candidate variants (fine-mapping) with different techniques. All the techniques and their applications are explained in the text.

### Replication phase

Replication of the findings in an independent study (either with a family-based or case-control design) is an essential step and will provide the most convincing evidence of a variant/gene causing the disease. Lack of replication may suggest a false-positive finding, but may also be due to genetic heterogeneity (described below) or a too small sample size.

*Allelic heterogeneity* This occurs when different mutations in the same gene lead to the same disorder. This type of heterogeneity



is often observed and it is therefore necessary to replicate the results. An entire gene/locus needs to be screened rather than just the single candidate variant. There are several examples of allelic heterogeneity underlying Mendelian and complex diseases, such as in ICF, a recessive immunoglobulin deficiency syndrome, which is caused by at least eleven different mutations in the same gene (40). MVID, an autosomal recessive disorder is caused by a wide spectrum of mutations in the *MYO5b* gene, including missense, nonsense, deletion and splice-sites mutations (26), and this thesis). Three rare mutations in the *NOD2* gene have been associated with Crohn's disease, possibly explaining association of the *NOD2* locus to this complex disorder (14). Mutations located in different regions of the gene (functional domains, binding sites, etc.) can have a different effect on the severity of the disease, and are observed as genotype-phenotype correlations (4, 12). For example, some genes influencing lipid levels carry common variants with a modest effect size leading to complex traits, as well as rare variants with a large effect size and causing Mendelian dyslipidemias (41, 42).

Locus heterogeneity This occurs when mutations in different genes cause the same disease. Lack of replication of previously mapped candidate regions can be due to high locus heterogeneity. For example, two different linkage regions were mapped for celiac disease in two large, independent, families of Dutch origin (29, 31). In the case of complex diseases, it was proposed to investigate extreme phenotypes of the disease that would be due to a more restricted group of loci and therefore easier to map and replicate (30). For Mendelian disorders, the locus heterogeneity is not as common as for complex diseases but it does occur on a regular basis. Griscelli syndrome, an autosomal recessive disease, is caused by a mutation in one of three interacting genes, *RAB27A-MLPH-MYO5A* a tripartite complex, leading to one of the three subtypes of this syndrome (43).

Clinical heterogeneity This is when a mutation in the same gene leads to different diseases. For example, mutations in the *RET*

gene can lead to two totally different diseases depending on their position in the gene: Hirschsprung disease that affects the colon, and familial medullary thyroid carcinoma (44). Classic examples of clinical heterogeneity are two types of muscular dystrophies caused by different deletions in the *DMD* gene. One is an out-of-frame deletion leading to Duchenne muscular dystrophy (mostly fatal before adulthood), whereas the second is an in-frame deletion leading to Becker muscular dystrophy, which is not life-threatening (45).

*Population heterogeneity* Each population differs in terms of allelic frequency, biological adaptation risk, and the prevalence of a disease. For severe Mendelian disorders, the mutated gene is more likely to be the same and mutations will occur with a very low frequency in each population as they undergo negative selection. Therefore, monogenic disorders are fairly replicable between different populations. The lack of replication of common SNPs associated to complex disorders in independent, case-control studies can be the result of genetic heterogeneity within a population due to **genetic drift**. As SNPs chosen from HapMap are fairly transferable between populations and create haplotypes due to LD, the **transferability** of a single SNP out of a group of correlated SNPs, rather than direct replication of the associated variant, might be a better way of replicating GWAS findings. This approach was successfully applied by Shiner et al. (2009), who replicated the results of association to a trait (adult height) in Europeans in an African population using two strategies: direct replication, in which only 8% of the loci were replicated, and locus-wide replication (transferability), in which 54% of the loci were replicated (46).

**Genetic drift:** is the change in the frequency of a variant (allele) in a population due to random sampling.

**Transferability:** is one of replication approaches in which all variants in high LD with the previously associated SNP (locus wide) are investigated for the association rather than single variants (direct replication).

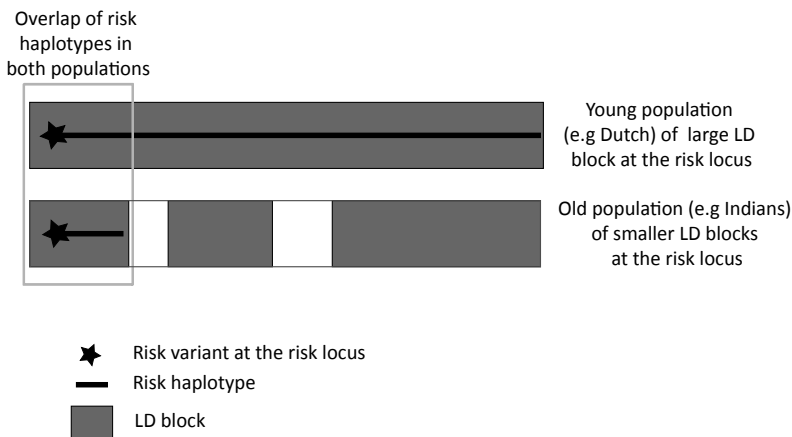
### Fine-mapping

Once the candidate region is mapped and successfully replicated, the next step is to fine-map it. Fine-mapping should narrow down the region or find the causative variant. The success of this step depends on the type of disease (complex, monogenic), the size of the region and its LD structure. The most common method is direct

re-sequencing of the region, by conventional Sanger sequencing or by next-generation, high-throughput sequencing, which has been successfully applied to fine-mapping of loci for monogenic diseases. Fine-mapping of GWAS loci for complex diseases is much more complicated as all variants from the associated loci are in rather high LD. Thus, cross-ethnic mapping is one of the alternatives to break up large haplotypes (47).

**Cross-ethnic fine-mapping** The concept behind this approach is to use unrelated populations to fine-map regions of high LD. Populations have different LD structures: older populations (Africans, Indians) are expected to have smaller LD blocks, whereas younger populations such as Europeans will have longer LD blocks (Figure 5).

In order to fine-map the region with this method the first step is to determine the risk haplotypes in both populations and to check whether they share a common ancestor and could possibly tag the same risk variant. This is crucial as some loci are associated with disease by more than one independent signal, suggesting more than one risk haplotype (10). The next step is to compare the LD block between the two populations at the locus of the risk haplotype and to narrow down the region based on the smallest LD block (Figure 5).



**Figure 5 Fine-mapping of the candidate region using two populations of different linkage disequilibrium (LD) structure.** In both populations the same loci were associated to the same diseases. The same or different causative variants are lying on the same haplotype. The haplotype in the older population is shorter due to the smaller LD blocks and can thus be used for fine-mapping the signal. The relevant narrower region is framed.

This approach was successfully used to fine-map the *IL2/IL21* region associated to celiac disease using Dutch and Indians populations (this thesis: Chapter 6). Such smaller haplotypes are more amenable for re-sequencing studies and the availability of haplotypes from different populations can assist in interpreting the results of such studies.

*Fine-mapping and re-sequencing by next-generation sequencing (NGS)*

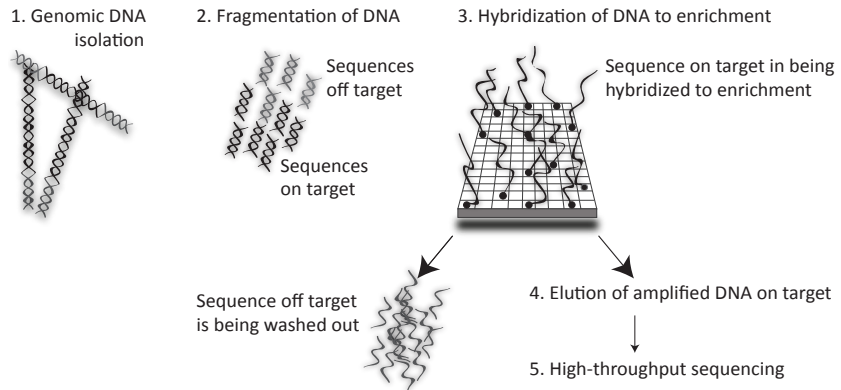
Next-generation sequencing (NGS) is used for high-throughput sequencing of the entire **exome** or genome of individuals for family-based or case-control studies. With the falling cost of NGS, it is rapidly becoming an alternative tool for the standard mapping of genetic diseases with problematic phenotypes (Mendelian or complex) in which GWAS simply does not tag the variant and a linkage method is not efficient because of *de novo* changes or locus heterogeneity (discussed earlier) (11). Since 2009, NGS has acquired high-throughput methods and become more affordable; a number of Mendelian disorders have been solved (48, 49) and causative variants have been found for some complex disorders (50). The difficulty of analyzing NGS data depends on the disease itself: looking for homozygous mutations causing recessive diseases is relatively easy compared to dominant disorders, as rare heterozygous sequence variants are rather common in the genome (51). Still, the analysis for all study designs and disease types can be divided into two steps: the first phase is a well-designed filtering method of good quality, well-covered SVs, and the second phase is checking for co-segregation or the statistical relevance of the finding. The first phase should lead to the discovery of one or more candidate variants. Filtering steps will vary depending on the disease model (e.g. dominant or recessive) and the study design (e.g. family-based or case-control design) and might include: (a) functionality of the SV, e.g. exon or coding regions, (b) the novelty or MAF of the SV, based on the publicly available data sets like **1000 Genomes project** (51), HapMap (36), or a private data set such as Genome of the Netherlands, (c) conservation of the region containing the SV, (d) zygosity of the SV, (e) function of the gene containing the SV and its likely involvement in the disease, etc. In a case-control design, the same filtering needs to be

**Exome:** is the genomic DNA representing all expressed messenger RNA sequences in any tissue.

**1000 Genomes Project:** the goal of this project is to find most of the genetic variants that have frequencies of at least 1% in the populations studied by using next-generation sequencing technology.

applied for a representative number of cases and controls (18, 20). If the number of variants after filtering is still high, other information can be added in order to filter out non-causative SVs. For example, Norton et al. (2011) applied expression data from the heart, assuming that a causative gene for dilated cardiomyopathy must be highly expressed in this organ (52). The second phase is to investigate the correlation between the sequence variants and disease by co-segregation of the selected SVs in families or testing for enrichment differences between cases and controls by applying appropriate statistical tests that pool together filtered SVs at a selected locus and testing them collectively (burden test) (53, 54). To prove the significance of a finding, we need to demonstrate replication by screening a large control panel and performing functional follow-up studies. NGS has been successfully applied to investigate a wide spectrum of diseases, such as autosomal dominant diseases (55), sporadic cases caused by *de novo* mutations (56), mitochondrial disease (57), X-linked diseases (58), autosomal recessive diseases (59), complex diseases with Mendelian-like segregation (50), and complex diseases in a case-control design (18, 19).

Nowadays NGS rather than Sanger-type sequencing is used for the efficient and fast screening of previously mapped regions. Candidate loci from linkage analysis are of the standard size of around 10-30 cM (= 10-30 Mb) and contain some 100-300 candidate genes (60). Loci associated in GWAS are smaller and contain fewer genes, but each of the complex diseases is associated to many such loci (61). The high number of genes and lack of a high-throughput screening method previously restricted the fine-mapping of candidate loci. But since NGS technology has become feasible and affordable, researchers have been able to screen for mutations in candidate regions by enriching for the entire exome and further analyzing regions of interest (55), or by achieving a very high coverage from enriching for selected regions of interest (62-64) (Figure 6).



**Figure 6 Targeted enrichment of genomic DNA.** Genomic DNA is isolated and its quality is checked (1). It is then sheared into small fragments (2) and selectively attached to the enrichment oligonucleotides (3) as the platform contains probes for regions of interest (on target). Sequences “off target” are washed off, whereas “on target” sequences are amplified and eluted (4) from the platform to be submitted for further sequencing (5).

In conclusion, NGS is proving to be a very powerful approach, which can be used as a screening method for hypothesis-driven as well as hypothesis-free approaches to investigate both Mendelian and complex diseases (Figure 4). With targeted re-sequencing (e.g. by exome enrichment), it is proving possible to solve most of the Mendelian disorders in a fast and efficient way. In the future, whole genome sequencing could be the best method for investigating complex diseases, as not only coding variants but regulatory and miRNAs can be re-sequenced. Both methods are being continually improved: exome enrichment capture is becoming more complete, containing more of the expressed regions, although whole genome technology yields a better coverage at the moment.

## References

1. Ingram G. I. (1976) The history of haemophilia. *J. Clin. Pathol.*, **29**, 469-479.
2. Ng S. B., Nickerson D. A., Bamshad M. J., Shendure J. (2010) Massively parallel sequencing and rare disease. *Hum. Mol. Genet.*, **19**, R119-R124.
3. Thomas P. D., Kejariwal A. (2004) Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for

- differences in molecular effects. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 15398-15403.
4. Botstein D., Risch N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.*, **33 Suppl**, 228-237.
  5. Garrod A. E. (1902) About Alkaptonuria. *Med. Chir Trans.*, **85**, 69-78.
  6. Mearin M. L., Ivarsson A., Dickey W. (2005) Coeliac disease: is it time for mass screening? *Best. Pract. Res. Clin. Gastroenterol.*, **19**, 441-452.
  7. Gibson G. (2009) Decanalization and the origin of complex disease. *Nat. Rev. Genet.*, **10**, 134-140.
  8. Schork N. J., Murray S. S., Frazer K. A., Topol E. J. (2009) Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.*, **19**, 212-219.
  9. Hindorff L. A., Sethupathy P., Junkins H. A., Ramos E. M., Mehta J. P., Collins F. S., Manolio T. A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 9362-9367.
  10. Trynka G., Hunt K. A., Bockett N. A., Romanos J., Mistry V., Szperl A., Bakker S. F., Bardella M. T., Bhaw-Rosun L., Castillejo G., *et al.* (2011) Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.*, **43**, 1193-1201.
  11. Van H., V, Yeyati P. L. (2004) Mechanisms of non-Mendelian inheritance in genetic disease. *Hum. Mol. Genet.*, **13 Spec No 2**, R225-R233.
  12. Manolio T. A., Collins F. S., Cox N. J., Goldstein D. B., Hindorff L. A., Hunter D. J., McCarthy M. I., Ramos E. M., Cardon L. R., Chakravarti A., *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747-753.
  13. Goldstein D. B. (2011) The importance of synthetic associations will only be resolved empirically. *PLoS. Biol.*, **9**, e1001008.
  14. Anderson C. A., Soranzo N., Zeggini E., Barrett J. C. (2011) Synthetic associations are unlikely to account for many common disease genome-wide association signals. *PLoS. Biol.*, **9**, e1000580.
  15. Cohen J. C., Kiss R. S., Pertsemlidis A., Marcel Y. L., McPherson R., Hobbs H. H. (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*, **305**, 869-872.
  16. Cohen J. C., Pertsemlidis A., Fahmi S., Esmail S., Vega G. L., Grundy S. M., Hobbs H. H. (2006) Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 1810-1815.
  17. Zimprich A., Benet-Pages A., Struhal W., Graf E., Eck S. H., Offman M. N., Haubenberger D., Spielberger S., Schulte E. C., Lichtner P., *et al.* (2011) A mutation in VPS35, encoding a subunit of the retromer complex, causes late-onset Parkinson disease. *Am. J. Hum. Genet.*, **89**, 168-175.
  18. Rivas M. A., Beaudoin M., Gardet A., Stevens C., Sharma Y., Zhang C. K., Boucher G., Ripke S., Ellinghaus D., Burt N., *et al.* (2011) Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.*, **43**, 1066-1073.
  19. Sulem P., Gudbjartsson D. F., Walters G. B., Helgadóttir H. T., Helgason A., Gudjonsson S. A., Zanon C., Besenbacher S., Bjornsdóttir G., Magnusson O. T., *et al.* (2011) Identification of low-frequency variants associated with gout and serum uric acid levels. *Nat. Genet.*, **43**, 1127-1130.
  20. Johansen C. T., Wang J., Lanktree M. B., Cao H., McIntyre A. D., Ban M. R., Martins R. A., Kennedy B. A., Hassell R. G., Visser M. E., *et al.* (2010) Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat. Genet.*, **42**, 684-687.
  21. Baker M. (2010) Genomics: The search for association. *Nature*, **467**, 1135-1138.
  22. Hardy J., Singleton A. (2009) Genomewide association studies and human disease. *N. Engl.*

- J. Med.*, **360**, 1759-1768.
23. Kell D. B., Oliver S. G. (2004) Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays*, **26**, 99-105.
  24. Sato T., Mushiake S., Kato Y., Sato K., Sato M., Takeda N., Ozono K., Miki K., Kubo Y., Tsuji A., *et al.* (2007) The Rab8 GTPase regulates apical protein localization in intestinal cells. *Nature*, **448**, 366-369.
  25. Muller T., Hess M. W., Schiefermeier N., Pfaller K., Ebner H. L., Heinz-Erian P., Ponstingl H., Partsch J., Rollinghoff B., Kohler H., *et al.* (2008) MYO5B mutations cause microvillus inclusion disease and disrupt epithelial cell polarity. *Nat. Genet.*, **40**, 1163-1165.
  26. Szperl A. M., Golachowska M. R., Bruinenberg M., Prekeris R., Thunnissen A. M., Karrenbeld A., Dijkstra G., Hoekstra D., Mercer D., Ksiazek J., *et al.* (2011) Functional characterization of mutations in the myosin Vb gene associated with microvillus inclusion disease. *J. Pediatr. Gastroenterol. Nutr.*, **52**, 307-313.
  27. Dawn T. M., Barrett J. H. (2005) Genetic linkage studies. *Lancet*, **366**, 1036-1044.
  28. van Belzen M. J., Vrolijk M. M., Meijer J. W., Crusius J. B., Pearson P. L., Sandkuijl L. A., Houwen R. H., Wijmenga C. (2004) A genomewide screen in a four-generation Dutch family with celiac disease: evidence for linkage to chromosomes 6 and 9. *Am. J. Gastroenterol.*, **99**, 466-471.
  29. Szperl A. M., Ricano-Ponce I., Li J. K., Deelen P., Kanterakis A., Plagnol V., van D. F., Westra H. J., Trynka G., Mulder C. J., *et al.* (2011) Exome sequencing in a family segregating for celiac disease. *Clin. Genet.*, **80**, 138-147.
  30. Dean M. (2003) Approaches to identify genes for complex human diseases: lessons from Mendelian disorders. *Hum. Mutat.*, **22**, 261-274.
  31. van Belzen M. J., Meijer J. W., Sandkuijl L. A., Bardoel A. F., Mulder C. J., Pearson P. L., Houwen R. H., Wijmenga C. (2003) A major non-HLA locus in celiac disease maps to chromosome 19. *Gastroenterology*, **125**, 1032-1041.
  32. Lander E. S., Botstein D. (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science*, **236**, 1567-1570.
  33. den Hollander A. I., Koenekoop R. K., Mohamed M. D., Arts H. H., Boldt K., Towns K. V., Sedmak T., Beer M., Nagel-Wolfrum K., McKibbin M., *et al.* (2007) Mutations in LCA5, encoding the ciliary protein lebercilin, cause Leber congenital amaurosis. *Nat. Genet.*, **39**, 889-895.
  34. Cordell H. J., Clayton D. G. (2005) Genetic association studies. *Lancet*, **366**, 1121-1131.
  35. Manolio T. A. (2010) Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.*, **363**, 166-176.
  36. Frazer K. A., Ballinger D. G., Cox D. R., Hinds D. A., Stuve L. L., Gibbs R. A., Belmont J. W., Bou-dreau A., Hardenbol P., Leal S. M., *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851-861.
  37. Franke A., McGovern D. P., Barrett J. C., Wang K., Radford-Smith G. L., Ahmad T., Lees C. W., Balschun T., Lee J., Roberts R., *et al.* (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.*, **42**, 1118-1125.
  38. Goldstein D. B., Cavalleri G. L. (2005) Genomics: understanding human diversity. *Nature*, **437**, 1241-1242.
  39. Wang K., Dickson S. P., Stolle C. A., Krantz I. D., Goldstein D. B., Hakonarson H. (2010) Interpretation of association signals and identification of causal variants from genome-wide association studies. *Am. J. Hum. Genet.*, **86**, 730-742.
  40. Wijmenga C., Hansen R. S., Gimelli G., Bjorck E. J., Davies E. G., Valentine D., Belohradsky B. H., van Dongen J. J., Smeets D. F., van den Heuvel L. P., *et al.* (2000) Genetic variation in ICF syndrome: evidence for genetic heterogeneity. *Hum. Mutat.*, **16**, 509-517.



41. Lusis A. J., Pajukanta P. (2008) A treasure trove for lipoprotein biology. *Nat. Genet.*, **40**, 129-130.
42. Kathiresan S., Willer C. J., Peloso G. M., Demissie S., Musunuru K., Schadt E. E., Kaplan L., Bennett D., Li Y., Tanaka T., *et al.* (2009) Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat. Genet.*, **41**, 56-65.
43. Van G. M., Dynoodt P., Lambert J. (2009) Griscelli syndrome: a model system to study vesicular trafficking. *Pigment Cell Melanoma Res.*, **22**, 268-282.
44. Hofstra R. M., Landsvater R. M., Ceccherini I., Stulp R. P., Stelwagen T., Luo Y., Pasini B., Hopfener J. W., van Amstel H. K., Romeo G., *et al.* (1994) A mutation in the RET proto-oncogene associated with multiple endocrine neoplasia type 2B and sporadic medullary thyroid carcinoma. *Nature*, **367**, 375-376.
45. Gillard E. F., Chamberlain J. S., Murphy E. G., Duff C. L., Smith B., Burghes A. H., Thompson M. W., Sutherland J., Oss I., Bodrug S. E., *et al.* (1989) Molecular and phenotypic analysis of patients with deletions within the deletion-rich region of the Duchenne muscular dystrophy (DMD) gene. *Am. J. Hum. Genet.*, **45**, 507-520.
46. Shriner D., Adeyemo A., Gerry N. P., Herbert A., Chen G., Doumatey A., Huang H., Zhou J., Christman M. F., Rotimi C. N. (2009) Transferability and fine-mapping of genome-wide associated loci for adult height across human populations. *PLoS. One.*, **4**, e8398.
47. Seldin M. F., Pasaniuc B., Price A. L. (2011) New approaches to disease mapping in admixed populations. *Nat. Rev. Genet.*, **12**, 523-528.
48. Ng S. B., Buckingham K. J., Lee C., Bigham A. W., Tabor H. K., Dent K. M., Huff C. D., Shannon P. T., Jabs E. W., Nickerson D. A., *et al.* (2010) Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.*, **42**, 30-35.
49. Ku C. S., Naidoo N., Pawitan Y. (2011) Revisiting Mendelian disorders through exome sequencing. *Hum. Genet.*, **129**, 351-370.
50. Musunuru K., Pirruccello J. P., Do R., Peloso G. M., Guiducci C., Sougnez C., Garimella K. V., Fisher S., Abreu J., Barry A. J., *et al.* (2010) Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. *N. Engl. J. Med.*, **363**, 2220-2227.
51. Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061-1073.
52. Norton N., Li D., Rieder M. J., Siegfried J. D., Rampersaud E., Zuchner S., Mangos S., Gonzalez-Quintana J., Wang L., McGee S., *et al.* (2011) Genome-wide studies of copy number variation and exome sequencing identify rare variants in BAG3 as a cause of dilated cardiomyopathy. *Am. J. Hum. Genet.*, **88**, 273-282.
53. Bansal V., Libiger O., Torkamani A., Schork N. J. (2010) Statistical analysis strategies for association studies involving rare variants. *Nat. Rev. Genet.*, **11**, 773-785.
54. Feng T., Elston R. C., Zhu X. (2011) Detecting rare and common variants for complex traits: sibpair and odds ratio weighted sum statistics (SPWSS, ORWSS). *Genet. Epidemiol.*, **35**, 398-409.
55. Nikopoulos K., Gilissen C., Hoischen A., van Nouhuys C. E., Boonstra F. N., Blokland E. A., Arts P., Wieskamp N., Strom T. M., Ayuso C., *et al.* (2010) Next-generation sequencing of a 40 Mb linkage interval reveals TSPAN12 mutations in patients with familial exudative vitreoretinopathy. *Am. J. Hum. Genet.*, **86**, 240-247.
56. Hoischen A., van Bon B. W., Gilissen C., Arts P., van L. B., Steehouwer M., de V. P., de R. R., Wieskamp N., Mortier G., *et al.* (2010) De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat. Genet.*, **42**, 483-485.
57. Haack T. B., Danhauser K., Haberberger B., Hoser J., Strecker V., Boehm D., Uziel G., Laman- tea E., Invernizzi F., Poulton J., *et al.* (2010) Exome sequencing identifies ACAD9 mutations as a cause of complex I deficiency. *Nat. Genet.*, **42**, 1131-1134.
58. Johnston J. J., Teer J. K., Cherukuri P. F., Hansen N. F., Loftus S. K., Chong K., Mullikin J. C.,

- Biesecker L. G. (2010) Massively parallel sequencing of exons on the X chromosome identifies RBM10 as the gene that causes a syndromic form of cleft palate. *Am. J. Hum. Genet.*, **86**, 743-748.
59. Bilguvar K., Ozturk A. K., Louvi A., Kwan K. Y., Choi M., Tatli B., Yalnizoglu D., Tuysuz B., Caglayan A. O., Gokben S., *et al.* (2010) Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature*, **467**, 207-210.
60. Glazier A. M., Nadeau J. H., Aitman T. J. (2002) Finding genes that underlie complex traits. *Science*, **298**, 2345-2349.
61. Park J. H., Wacholder S., Gail M. H., Peters U., Jacobs K. B., Chanock S. J., Chatterjee N. (2010) Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.*, **42**, 570-575.
62. Nejentsev S., Walker N., Riches D., Egholm M., Todd J. A. (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, **324**, 387-389.
63. Hopp K., Heyer C. M., Hommerding C. J., Henke S. A., Sundsbak J. L., Patel S., Patel P., Consugar M. B., Czarnecki P. G., Gliem T. J., *et al.* (2011) B9D1 is revealed as a novel Meckel syndrome (MKS) gene by targeted exon-enriched next-generation sequencing and deletion analysis. *Hum. Mol. Genet.*, **20**, 2524-2534.
64. Otto E. A., Hurd T. W., Airik R., Chaki M., Zhou W., Stoetzel C., Patil S. B., Levy S., Ghosh A. K., Murga-Zamalloa C. A., *et al.* (2010) Candidate exome capture identifies mutation of SDC-CAG8 as the cause of a retinal-renal ciliopathy. *Nat. Genet.*, **42**, 840-850.



***Functional characterization of mutations in the myosin Vb gene associated with microvillus inclusion disease***

*Agata Szperl<sup>1,\*</sup>, Magdalena R. Golachowska<sup>2,\*</sup>, Marcel Bruinenberg<sup>1</sup>, Rytis Prekeris<sup>3</sup>, Andy-Mark W. H. Thunnissen<sup>4</sup>, Arend Karrenbeld<sup>5</sup>, Gerard Dijkstra<sup>6</sup>, Dick Hoekstra<sup>2</sup>, David Mercer<sup>7</sup>, Janusz Ksiazek<sup>8</sup>, Cisca Wijmenga<sup>1</sup>, Martin C. Wapenaar<sup>1</sup>, Edmond H. H. M. Rings<sup>9,#</sup>, Sven C. D. van IJzendoorn<sup>2,#</sup>.*

*\* # These authors contributed equally.*

*Departments of <sup>1</sup>Genetics, <sup>2</sup>Cell Biology/Membrane Cell Biology, <sup>5</sup>Pathology, <sup>6</sup>Gastroenterology and Hepatology, and <sup>9</sup>Pediatrics, University Medical Center Groningen and University of Groningen, Groningen, the Netherlands*

*<sup>3</sup>Department of Cellular and Developmental Biology, School of Medicine, University of Colorado Health Sciences Centre, Denver, Colorado, USA*

*<sup>4</sup>Department of Biophysical Chemistry, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, The Netherlands*

*<sup>7</sup>University of Nebraska Medical Center, Omaha, NE/USA; <sup>8</sup>Department of Pediatrics, Children's Memorial Health Institute, Warsaw, Poland*

*This manuscript was published in Journal of Pediatric Gastroenterology and Nutrition, 2011, 52(3):307-13.*

## Abstract

**Objectives.** Microvillus inclusion disease (MVID) is a rare autosomal recessive enteropathy characterized by intractable diarrhea and malabsorption. Recently, various *MYO5B* gene mutations have been identified in MVID patients. Interestingly, several MVID patients showed only a *MYO5B* mutation in one allele (heterozygous) or no mutations in the *MYO5B* gene, illustrating the need to further functionally characterize the cell biological effects of the *MYO5B* mutations.

**Methods.** The genomic DNA of nine patients diagnosed with microvillus inclusion disease was screened for *MYO5B* mutations, and qPCR and immunohistochemistry on the material of two patients was performed to investigate resultant cellular consequences.

**Results.** We demonstrate for the first time that *MYO5B* mutations can be correlated with altered myosin Vb mRNA expression and with an aberrant subcellular distribution of the myosin Vb protein. Moreover, we demonstrate that the typical and myosin Vb-controlled accumulation of rab11a- and FIP5-positive recycling endosomes in the apical cytoplasm of the cells is abolished in MVID enterocytes, which is indicative for altered myosin Vb function. Also, we report 8 novel *MYO5B* mutations in 9 MVID patients of various ethnic backgrounds, including compound heterozygous mutations.

**Conclusions.** Our functional analysis indicate that *MYO5B* mutations can be correlated with an aberrant subcellular distribution of the myosin Vb protein and apical recycling endosomes which, together with the additional compound heterozygous mutations, significantly strengthen the link between *MYO5B* and MVID.

## Introduction

A structurally, compositionally, and functionally distinct plasma membrane at the apex of the intestinal epithelial cell monolayer provides a selective and protective barrier that regulates the uptake of nutrients from the lumen. The inability of intestinal cells to maintain an apical brush border and the consequence it has on human functioning becomes particularly apparent in patients diagnosed with microvillus inclusion disease (MVID; OMIM 251850)

Microvillus inclusion disease is a rare autosomal recessive disease presenting with severe intractable diarrhea and malabsorption in neonates (1-5). At the cellular level, variable brush border atrophy with accumulation of lysosomal granules and microvillus inclusions is observed in the apical cytoplasm of MVID enterocytes (1, 2, 6, 7). Periodic acid-Schiff-stained and other apical brush border components (e.g. CD10) are typically absent from the cell surface and accumulate in compartments in the apical cytoplasm (8). In contrast to the apical proteins, basolateral proteins display a normal polarized distribution at the surface of MVID enterocytes (8, 9) which appear normally arranged in monolayers with distinguishable cell-cell adhesion junctions.

MVID is often described in children born of consanguineous parents and this allowed Müller et al. (10) to map the MVID locus to 18q21 using homozygosity mapping in an extended Turkish kindred. Mutation analysis of a positional candidate gene from the region of homozygosity, *MYO5B*, revealed a homozygous in-frame insertion in the MVID patients from the Turkish kindred. To date, 25 different nonsense, missense, splice site, or in-frame insertion mutations in the *MYO5B* gene (OMIM# 606540) have been identified in 28 MVID patients from consanguineous and unrelated marriages (10-12). The *MYO5B* gene encodes myosin Vb, which is an actin filament-based motor protein that interacts with and regulates among others the subcellular spatial distribution of recycling endosomes that express small GTPase proteins such as Rab11a on their cytoplasmic surface.

In several MVID patients *MYO5B* mutations were found in only one allele (heterozygous) or no *MYO5B* mutation was found (10). Moreover, although knockdown of myosin Vb in human epithelial colorectal adenocarcinoma (Caco-2) cells recapitulates most of the cellular phenotypes of MVID (12), it is not known whether myosin Vb mRNA and protein expression and myosin Vb function is affected in MVID patients. Such information would support *MYO5B* gene screening as a

diagnostic tool for this difficult to recognize rare disease, and allow reliable genetic counseling and prenatal screening. Supporting evidence that *MYO5B* mutations have consequences for the expression and/or function of the myosin Vb protein in MVID enterocytes as well as mutational analyses of additional MVID patients are therefore imperative. In this study we have used small intestine biopsies to demonstrate that MVID-associated *MYO5B* mutations affect the expression and function of the myosin Vb protein in MVID enterocytes. In addition, we have performed mutation analyses of 9 additional MVID patients of various ethnic background and report 8 new *MYO5B* mutations; three homozygous and five heterozygous mutations which include stop codons/nonsense mutations, missense mutations, splice site mutations, large deletions, and compound heterozygous mutations.

## **Materials and methods**

### **Description of patients and clinical history**

Nine patients in whom histological examination of small intestine mucosa confirmed the diagnosis of MVID were included in this study. Patients 1-6 were collected from a larger patient cohort that received a bowel transplant via the Liver/Small Bowel Transplant Program of the University Of Nebraska Medical Center (USA). Patient 1 is a 1 year old Hispanic male with early-onset MVID from reported non-consanguineous parents. His brother died of MVID at 21 months of age. Patient 2 is a 3 year old Hispanic female with early-onset MVID of consanguineous parents (first cousins). She has two sisters with MVID one of which is patient 3. Patient 3 is a 5 year old Hispanic female with early-onset MVID from consanguineous parents (first cousins). She has two sisters with MVID one of which is patient 2. Patient 4 is a 1 year old Navajo Indian male with early-onset MVID of related parents (died of sepsis). Patient 5 is a 12 year old Navajo Indian female with early-onset MVID from related parents (died of sepsis with multi-organ system failure). She has two healthy sisters. Patient 6 is a 0 year old Caucasian female with early onset MVID from reported non-consanguineous parents (died of sepsis from aspergillus and continuing acute rejection). She has one healthy sibling and two siblings died with unknown cause. Patient 7 is a 1 year old Polish-Caucasian female with early-onset MVID from reported non-consanguineous parents. Patient 8 was a 5 year old Moroccan boy with early-onset MVID from consanguineous parents (first-degree cousins). Patient 9 is a 5 year old Dutch-Caucasian boy from unrelated parents who

was diagnosed with late-onset MVID. Unaffected parents and siblings of patient 7, 8 and 9 were also recruited and, after informed consent, saliva samples were collected and genomic DNA was extracted. In addition, available duodenal tissue from patient 8 and 9 and age-matched normal control patients was obtained and processed for immunohistochemistry. Also two 2 year old Dutch girls (twin) of non-consanguineous parents who presented severe secretory diarrhea and nutrient malabsorption directly after birth, but did not display the diagnostic light and electron microscopical hallmarks of MVID, were included in the study (patients 10 and 11). Written consent was obtained for all patients. This study has been reviewed and approved by the University Medical Center Groningen review board.

### **DNA and RNA isolation**

DNA was isolated from peripheral blood samples using standard laboratory procedures. DNA and RNA from saliva was also collected and isolated (OrageneDNA and OrageneRNA, DNA Genotek Inc, Ottawa, Canada). RNA from in liquid nitrogen snap-frozen biopsy samples was isolated after homogenization using 1 mm glass beads using Trizol (Invitrogen, Carlsbed, CA). Concentration and purity were determined with NanoDrop ND-1000 (Isogen Life Science, De Meern, the Netherlands).

### **RT-PCR**

Real-time PCR reaction was performed for the quantification of *MYO5B*. RNA was isolated from duodenum biopsies of twelve controls and patients 8 and 9. cDNA was generated with a High Capacity cDNA Archive kit (Applied Biosystems, Foster City, CA) using 1µg total RNA. Primers were designed with Primer Express v.3 (Applied Biosystems); RT\_*MYO5B*for: TTGGAAGTGTGGCGATTGAG; RT\_*MYO5B*Brev: GCAGTCGGCAGAAGTTGCTT. For *GUSB* expression, we used a TaqMan Pre-Developed Assay (Applied Biosystems). Reactions consisted of 1xSYBR Green (or Universal) PCR Mastermix, 1mM of each primer and 1µl cDNA. Cycling conditions were 50°C for 2 min, 95°C for 10 min and 40 cycles of 95°C for 15 s and 60°C for 1 min. Results were analyzed using SDS v.2.3 (Applied Biosystems).



## Sequencing

The *MYO5B* coding region and splice sites were PCR amplified and directly sequenced in probands. Their relatives and approximately 50 control individuals (~100 chromosomes) were screened for the detected mutations. Primers for PCR amplification (supplementary tables 1 and 2 show primers used for amplification of genomic DNA and cDNA, respectively) were designed using *Primer3*<sup>21</sup> on the genomic sequence of *MYO5B* (NC\_000018.8) and its mRNA (NM\_001080467). The PCR reaction was performed with 50ng genomic DNA in 20ul reaction volume which included 1xPCR buffer-A (GE Healthcare, Piscataway/NJ), 2.5 mM dNTPs, 1mM primers (Eurogentec, San Diego, CA), 0.5U Taq polymerase (GE Healthcare). For exons 1, 2, 12, 17, and 18, PCR reaction was performed with 150ng of genomic DNA in 25ul reaction volume with 1xPCR buffer (Buffer-B) made of 0.1M Tris-HCL (pH8.8), 0.1M MgCl<sub>2</sub>, 0.01M mercaptoethanol, 0.05M ethylenediaminetetraacetic acid/0.1M (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub>. The PCR conditions differed with respect to the annealing temperatures and buffers used. Initial denaturation at 95°C for 5 min (4 min Buffer-B); 40 cycles (33 cycles Buffer-B) of denaturation at 95°C for 30 s (1 min Buffer-B), annealing for 30 s (1 min Buffer-B), and extension at 72°C for 30 s (2 min Buffer-B); final extension at 72°C for 5 min (7 min Buffer-B). PCR products were purified (37°C for 15 min, 80°C for 15 min) with ExoSap-IT (USB, Cleveland, Ohio) and Sephadex columns. Sequencing reactions were performed using BigDye terminator mix (Applied Biosystems). Sequences were read on a 3730 DNA analyzer and 3130 Genetic analyzer (Applied Biosystems) and we aligned sequencing data with control and reference sequences using ContigExpress software (Invitrogen, Carlsbed, CA).

## Deletion detection

In order to detect large deletions in *MYO5B* in patient 7, real-time quantitative PCR was used to determined copy numbers of the exons. Reactions consisted of 1xSYBR Green PCR Mastermix (Applied Biosystems), 1mM of each primer (supplementary table 2) and 25ng of genomic DNA. Cycling conditions were 50°C for 2 min, 95°C for 10 min and 40 cycles of denaturation at 95°C for 15s, and annealing for 1 min.

## Immunohistochemistry

Duodenal biopsies of MVID patients 8 and 9 and age-matched controls were

fixed in paraffin, and cut in 3  $\mu\text{m}$  thick sections. Slides were dried overnight in 60°C and deparaffinized in xylol-100%-96%-70%ethanol and demiwater. Epitopes were retrieved by protease digestion or in citric acid pH6.0 (autoclaved; 5 min, 120°C). Endogenous peroxidase was deactivated with 3.5%  $\text{H}_2\text{O}_2$ . Following blocking of non-specific binding sites in 4% normal-goat-serum, slides were incubated with primary antibodies, washed, and incubated with appropriate horseradish peroxidase-conjugated secondary antibodies. Diaminobenzidine was used as a substrate for peroxidase. Hematoxyline was used to stain the nuclei. Slides were dehydrated with ethanol, dried and mounted. Antibodies used: polyclonal antibodies raised against a synthetic peptide derived from the C-terminal hypervariable region of the human Rab11a sequence (Zymed Laboratories Inc); polyclonal antibodies against Rip11/FIP5 (13); polyclonal antibodies raised against a synthetic peptide corresponding to C- or N-terminal residues (amino acids 1093-1112 or 23-41, respectively) of human myosin Vb (Antagene Inc; 60B923) that recognizes a single band of the appropriate molecular mass of ~214 kDa on Western blot; Horseradish peroxidase-conjugated donkey anti-rabbit, sheep anti-mouse antibodies (GE Healthcare). Immunohistochemistry images show villus cells.

## Results

### **Eight new *MYO5B* mutations associated with nine microvillus inclusion disease patients**

*MYO5B* is composed out of 40 coding exons which were separately amplified and subjected to sequence analysis. All eleven patients were included in the mutation analysis by direct sequencing of the entire gene in both forward and reverse directions. Six patients revealed homozygous mutations. Patient 6 revealed one heterozygous change, while patient 7 and 9 carry compound heterozygous mutations (Table 1). Patients 10 and 11, who presented nutrient malabsorption and intractable secretory diarrhea after birth but were not diagnosed with MVID, did not reveal *MYO5B* mutations.

Patient 1 carries a homozygous non-conservative missense mutation in exon 8 (c.946G>A, p.Gly316Arg), which replaces a small aliphatic glycine (conserved in myosin Va and Vc; Supplementary figure 1) with a large and charged arginine in the protein's conserved head domain region. In patients 2 and 3 we found a shared homozygous deletion in exon 19 (c.2330\_del G; Supplementary figure 2).

This mutation disturbs the reading frame and leads to a premature stop codon (p.Gly777AsnfsX6; Supplementary figure 3) in the first calmodulin-binding IQ1 motif of myosin Vb. Any resultant protein will therefore not be able to dimerize and function as a processive motor protein, and lacks the entire cargo-binding tail domain. Patients 4 and 5 are homozygous for a non-conservative missense mutation in exon 16 (c.1979C>T, p.Pro660Leu). This mutation was recently described (11) in 7 Navajo MVID patients. In patient 6 we found one *cx* heterozygous mutation in exon 19 which results in a premature stop codon (c.2246C>T, p.Arg749X) in the head domain of myosin Vb (p.Arg749 is conserved in myosin Va and Vc; Supplementary figure 4). Resultant protein will not be able to dimerize and function as a processive motor protein, and lack the entire cargo-binding tail domain.

For patient 7, 8 and 9, we also obtained DNA samples from unaffected siblings and/or parents. Patient 7 reveals a compound heterozygous mutation, which includes a paternal allele with a non-conservative asparagine-to-serine (c.1367A>G, p.Asn456Ser) substitution in exon 11 of the head domain (Figure 1A, B, D) (conserved in myosin Va and Vc; Supplementary figure 5), together with a missense variant p.Met1688Val (c.5062A>G) in exon 37 (p.Met1688 is substituted in *MYO5A* and *MYO5C*; Supplementary figure 6). p.Met1688Val represents an infrequent polymorphism, as it was found in Polish and Dutch controls with allele frequencies of 5.8% (6/104) and 1.7% (2/116), respectively. When searching for a maternally transmitted mutation in patient 7, a Mendelian inconsistency in the inheritance of the exon11 variant c.1367A>G was observed: the mother appeared to be homozygous A/A while the patients was homozygous G/G (Figure 1A). This could possibly point towards a maternal transmission of a deletion. Using real-time PCR to determine the copy number of the *MYO5b* gene, we found that the maternal allele in patient 7 contained a deletion involving exons 2–12 of *MYO5B* (Figure 1C), rendering any protein formed incapable of binding actin and function as a motor protein. Sequencing of *MYO5B* in patient 8 revealed a homozygous stop codon in exon 33 (c.4366C>T, p.Gln1456X) (Figure 1A) which removes the terminal Rab11a-binding sites (1799-1814) (Figure 1B), which only functions in unison with the more proximal Rab11a-binding site (1400-1415). Sequencing of *MYO5B* in patient 9 showed that this patient is a compound heterozygote carrying a *de novo* non-conservative substitution mutation in exon 12 (c.1540T>C, p.Cys514Arg), and a maternally derived mutation in intron 33 (c.4460-1G>C) that destroys the canonical

splice acceptor (SA) site (Figure 1A, B). Intron 33 harbors three clusters of potent candidate cryptic SA-sites (Supplementary figure 7). PCR on the patient's intestinal cDNA with primers for intron 33 and exon 35 demonstrated retention of >100 bp of intron 33 immediately upstream of exon 34. This 'extended exon 34' contains nine stop codons, at least one in each of the three reading frames (Supplementary figure 8). The p.Cys514 residue forms part of the helix-turn-helix motif in the motor

**Table 1. Summary of MYO5B mutations associated with MVID as reported in this study.** NMD, nonsense-mediated RNA decay; PMT, prematurely terminated protein; Rab11a BD (binding domain); n.r., not reported.

Subject	Ancestry	Parental Consang.	Sex/onset	MYO5B mutation	Homo/heterozygous	MYO5b domain	Predicted effect on RNA and/or protein
1	Hispanic	n.r.	Male/early	946 G>A (G316R, exon 8)	Homozygous	head	Nonconservative substitution, evolutionary conserved
2, 3	Hispanic	yes	Female/early	2330 del G (out of frame, leads to stop codon, exon 19)	Homozygous	Neck (IQ1)	Nonconservative substitution, evolutionary conserved, PMT, loss of dimerization and cargo-binding domains
4, 5	Navajo Indian	yes	Male/early; Female/early	1979C>T (P660L, exon 16)	Homozygous	Head	Nonconservative substitution, evolutionary conserved
6	Caucasian	No	Female/Early	c.2246 C>T (R749X, exon 19)	Compound Heterozygous	Neck (IQ1)	Nonconservative substitution, evolutionary conserved, PMT, loss of dimerization and cargo-binding domains
7	Polish	No	Female/early	DIV51_IV512 (deletion exon 2-12)	Compound Heterozygous	Head	Shortened protein, in frame deletion residues 10-515
7	Polish	No	Female/early	1367A>G (N456S; exon 11)	Compound Heterozygous	Head	Nonconservative substitution, evolutionary conserved
8	Dutch	No	Male/late	1540T>C (C514R, exon 12)	Compound Heterozygous	Head	Nonconservative substitution, evolutionary conserved
8	Dutch	No	Male/late	IV533+3753G>C (splicing, intron 33)	Compound Heterozygous	Tail	Partial intron 33 insertion, NMD; PMT, loss of distal Rab11a BD
9	Moroccan	Yes	Male/early	4366C>T (Q1456X, exon 33)	homozygous	Tail	NMD; PMT, loss of distal Rab11a BD

domain that is associated with actin-binding (14-17) (Figure 1D). All mutations identified in this study are listed in Table 1. The position of all mutated residues in the crystal structure of the myosin Vb head domain are depicted in Figure 1D.

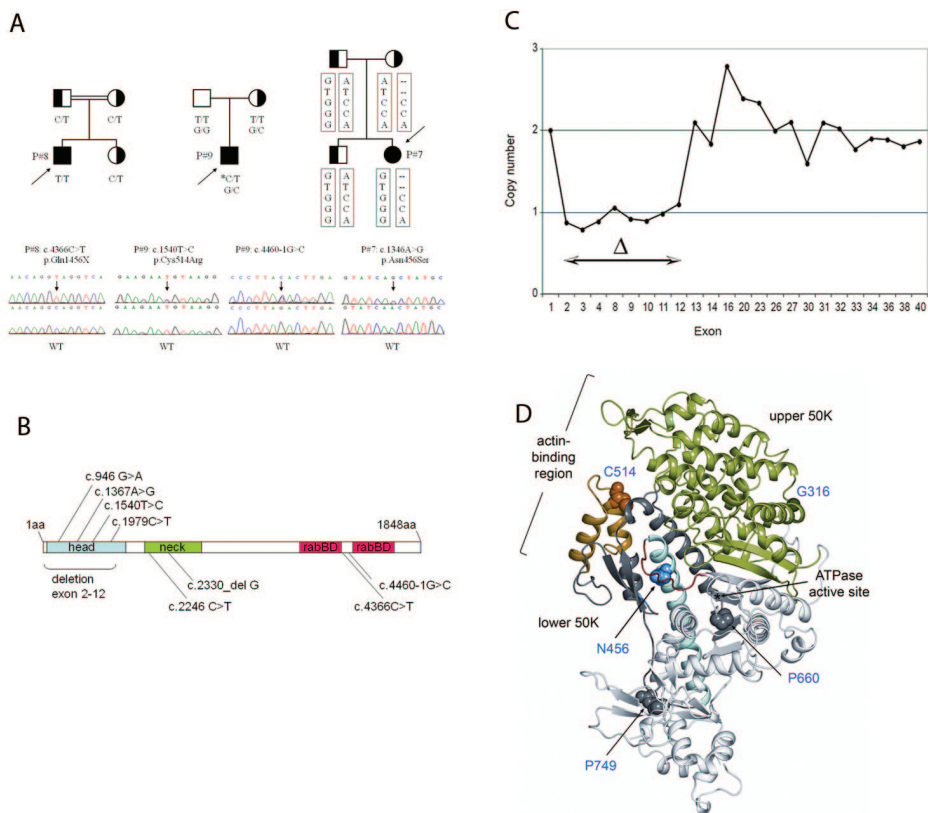
Resequencing mutation-containing exons revealed that none of the identified mutations were detected in 50 ethnically matched controls, or have been reported as known variants (in HapMap, dbSNP, and the 1000 genome database), unless stated otherwise.

### ***MYO5B* mutations affect the expression and function of the myosin Vb protein in MVID enterocytes**

We analyzed the expression levels of myosin Vb mRNA from biopsies of patients 8 and 9 by real-time PCR and compared these to control patients. In patient 8, myosin Vb mRNA expression was reduced by 50% when compare to 14 non MVID control patients. (Figure 2A), which is in agreement with the identified nonsense mutation p.Gln1456X which is predicted to result in nonsense-mediated RNA decay (18). In patient 9 myosin Vb mRNA levels were comparable to controls (Figure 2A).

We also analyzed the cellular expression pattern of myosin Vb protein in duodenal biopsies of patients 8 and 9 and age-matched controls. The myosin Vb protein is present in the villus enterocytes and mainly concentrated at their apical aspect below the brush border of control enterocytes (Figure 2B, arrow). In contrast, no or little specific myosin Vb signal was detected in the enterocytes of patient 8 and 9, respectively (Figure 2B). The *MYO5B* mutations did not involve residues that were used in the synthetic peptides to generate these antibodies, and the antibodies should recognize the mutant protein if present. The lack of clear myosin Vb signal in MVID enterocytes may reflect the absence of the protein (in accordance with the reduced myosin Vb mRNA levels in patients 8; see above), and/or may reflect a dispersion of remaining myosin Vb protein throughout the cells rendering myosin Vb below the detection limit.

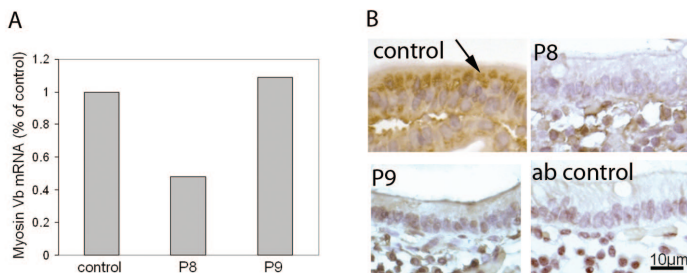
Myosin Vb regulates the subcellular positioning of recycling endosomes by binding to small GTPase Rab proteins such as Rab11a at the cytosolic surface of endosomes and attaching these endosomes to and moving them along actin filaments (19-24). Alterations in the typical spatial organization of recycling endosomes in MVID enterocytes can therefore be used as a read-out for altered myosin Vb function. To address this, the expression and distribution of recycling



**Figure 1. Identification and mutation analysis of *MYO5B* in MVID patients and parents/siblings.**

A) Pedigrees in three families with MVID patients 7, 8 and 9 (designated P7, P8, and P9). The c.1540T>C substitution in patient 9 is a *de novo* mutation on the paternal chromosome (indicated with asterisk). Haplotype analysis in patient 7 and her family was indicative for a deletion in *MYO5B* on the maternal chromosome. The *MYO5B* variants tested were (top to bottom): c.1367A>G (exon 11), c.3276+11 (rs2276176, intron 24), c.4222-73 (rs490648, intron 31), c.4315+5 (rs488890 intron 32), c.5062A>G (exon 37), c.5313+72(rs621101, intron38)

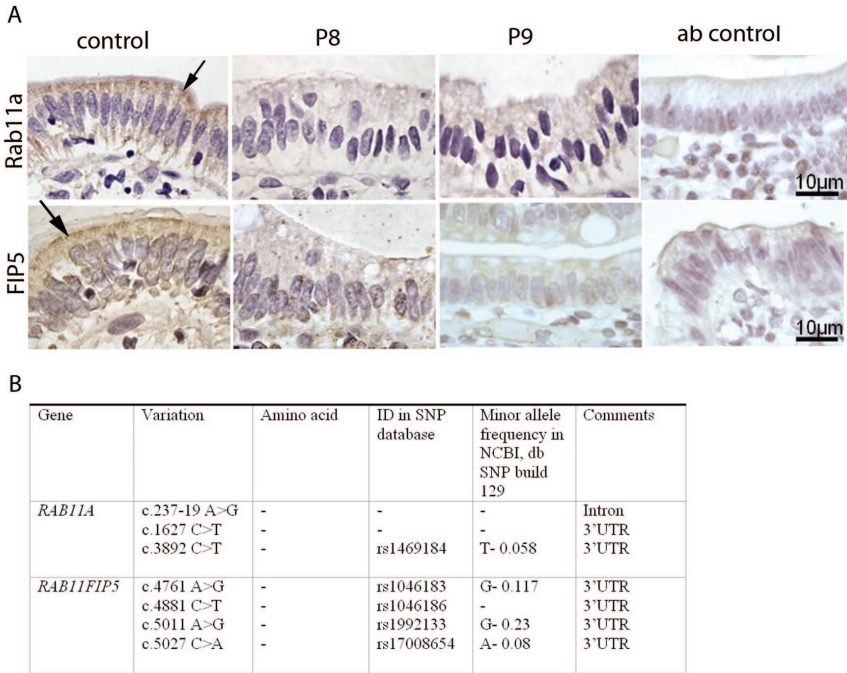
B) *MYO5B* mutations identified in MVID patients. C) Deletion mapping results of patient 7. The maternally derived deletion, spanning exons 2–12, was determined by quantitative PCR from the proband's brother (2 copies for each exon) to normalize the signal. D) Ribbon diagram of the nucleotide-free structure of the motor domain of chicken myosin V (rigor-like/strong actin-binding state, PDB code: 1OE9), indicating the locations of mutated residues (figure prepared with PyMOL (DeLano Scientific LLC)).



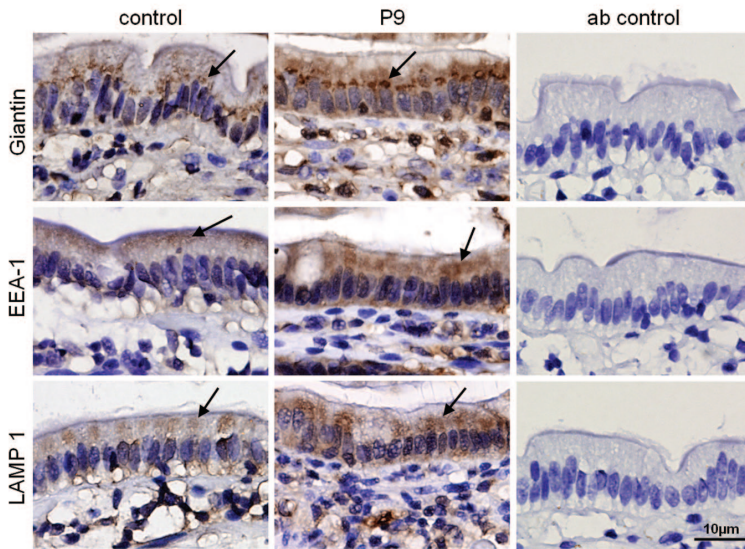
**Figure 2. Real-time PCR detection of myosin Vb mRNA and immunohistochemical labeling of myosin Vb.**

A) RNA was extracted from small intestinal biopsies of MVID patient 8 and 9 (designated P8 and P9) as described in Materials and methods, and relative myosin

Vb mRNA levels was determined with real-time PCR. B) small intestinal biopsies of MVID patient 8 and 9 and age-matched controls were labeled with antibodies against human myosin Vb. Negative (non-immune first antibody) control staining are shown. The accumulation of myosin Vb in the apical cytoplasm in control enterocytes (arrows) is lost in MVID enterocytes.



**Figure 3. Immunohistochemical labeling and variant analysis of Rab11a and FIP5(/Rip11).** A) The distribution of Rab11a and FIP5(/Rip11) in control and MVID enterocytes of patients 8 and 9 (designated P8 and P9) is shown. Negative (non-immune first antibody) staining is shown (Ab control). Note that the accumulation of Rab11a in the apical cytoplasm of control enterocytes (arrows) is lost in MVID enterocytes. B) Sequence variants in the genes *RAB11A* and *RAB11FIP5*.



**Figure 4. Immunohistochemical labeling of early endosome-, late endosome/lysosome, and Golgi-associated proteins.** The subcellular distributions of giantin, EEA1 and LAMP-1 in control and MVID enterocytes (patient 9 (P9)) are shown. Negative (non-immune first antibody) control staining are shown (ab control).

endosome-associated proteins Rab11a (19-24) and the Rab11a effector protein FIP5 (Rip11) (13) was investigated. The recycling endosome-associated proteins Rab11a and FIP5 (also known as Rip11) accumulate just below the enterocyte brush border close to the apical membrane in control duodenal tissue, similar to myosin Vb (Figure 3A, arrows). In contrast, in MVID enterocytes, Rab11a and FIP5 did not accumulate in the apical region and, instead, no specific staining pattern (compare to negative antibody (ab) control), could be observed (Figure 3A). Sequence analysis revealed single nucleotide polymorphisms but no functional mutations in the coding regions of the genes *RAB11A* and *RAB11FIP5* in patients 8, and 9 (summarized in Figure 3B). Early endosomal antigen 1 (EEA1), a marker of early sorting endosomes and typically excluded from myosin Vb-positive recycling endosomes, and the late endosome- and lysosome-associated protein LAMP-1 displayed comparable staining patterns in controls and MVID patient 9 (Figure 4, arrows). The distribution of the Golgi complex was also apparently unaffected in MVID enterocytes, although it appeared somewhat more concentrated in the supranuclear region (Figure 4).

## Discussion

We have analyzed the sequence of *MYO5B* in 9 MVID patients and identified 8 new mutations (~25% of all reported mutations) including a large deletion, a single nucleotide deletion, two missense mutations, and one nonsense mutation. We also report two additional compound heterozygous *MYO5B* mutations. In two patients we found a homozygous missense mutation that has been described previously (11). Our study adds 8 mutations to the 25 earlier reported by Müller and colleagues (24 mutations/21 patients) (10, 12) and Erickson and colleagues (one mutation shared by 7 Navajo patients) (11), yielding a total of 33 distinct *MYO5B* mutations in 37 MVID patients that have been identified to date. With our data, providing 25% of all currently reported *MYO5B* mutations and patients, we make a first analysis of the current *MYO5B* mutation spectrum. Of the 33 thus far published *MYO5B* mutations, 24 are localized in the N-terminal head domain that includes actin-binding and ATP catalytic sites, 2 in calmodulin-binding IQ motifs that form the light chain-binding lever arm domain, 1 in a potential coiled-coil regions that mediates the association of the heavy chain into dimers, and 6 are localized in the cargo-binding globular tail domain. All of the *MYO5B* mutations are distinct from those reported in *MYO5A* and other nonconventional myosins. Furthermore, the



reported heterozygous mutations are exclusively found in patients of Caucasian origin (Polish, Irish, French, and USA), and include at least one nonsense mutation or large deletion. Interestingly, all but one missense mutations cluster in the myosin Vb head domain, whereas the nonsense, splice-site and deletions/insertions are found randomly in the motor, lever arm, and tail domain. While some of the mutations are predicted to result in nonsense-mediated RNA decay (e.g. the homozygous p.Gln1456X mutation in patient 8 and the c.4460-1G>C mutation in patient 9), which is supported by the observed reduction in myosin Vb mRNA levels in patient 8, other *MYO5B* mutations involve residues that are important for the function of the myosin Vb protein. Indeed, the N456 residue mutated in patient 7, for instance, is part of a set of conserved motifs shared in all myosins that participate in coupling changes in the ATPase active site (P-loop and switch I) to conformational changes in the actin-binding and force-generating domains, and proposed to have a pivotal role in motor function as mediator of allosteric communication (14-17).

We demonstrate that a nonsense *MYO5B* mutation correlates with reduced myosin Vb mRNA expression, and that *MYO5B* mutations correlate with an aberrant cellular expression pattern of the myosin Vb protein. A main function of myosin Vb is to regulate the subcellular distribution and positioning of recycling endosomes. It does so by interacting with small GTPase Rab proteins such as Rab11a at the cytosolic surface of recycling endosomes and coupling these endosomes to and positioning them along actin filaments. We demonstrate that the typical and myosin Vb-controlled accumulation of Rab11a- and FIP5-positive recycling endosomes in the apical cytoplasm of the cells is abolished in MVID enterocytes. These data are indicative for an altered myosin Vb function in MVID enterocytes. It should be noted that our conclusions are based on two cases of this rare disease and that future experiments are necessary to further consolidate these. Further in-depth analysis of *MYO5B* mutations and their molecular and cell biological consequences are warranted to expand our understanding of how they are related to MVID pathogenesis.

It is encouraging that all MVID patients (except for one (10)) that have been screened thus far carry mutations in their *MYO5B* gene, and the discovery of additional compound heterozygous mutations by Ruemmele and colleagues (12) and us (this study) significantly strengthens the correlation between *MYO5B* and MVID. This correlation is further supported by a recent study (12) in which

knockdown of myosin Vb in human epithelial colorectal adenocarcinoma (Caco-2) cells recapitulates most of the cellular phenotypes of MVID, and by our observation that *MYO5B* mutations were not found in two patients that presented with secretory diarrhea and malabsorption after birth but were not diagnosed with MVID. A firm association of *MYO5B* mutations with MVID is a major advance in the diagnosis of this rare but fatal disease, in which variable phenotypes are seen among patients. It will also facilitate reliable genetic counseling and prenatal screening. Because total parenteral nutrition and bowel transplants are, at best, non-permanent solutions for treating this devastating disease, the continuing identification of *MYO5B* mutations will pave the way for the development of alternative therapeutic strategies.

### Acknowledgements

We thank the patients, their parents and siblings, and the transplantation teams of the UMC Groningen and the Children's Memorial Health Institute in Warsaw. We thank Carolien Gijsbers (Juliana Children's Hospital, The Hague, The Netherlands) and Marc Benninga (Academic Medical Center Amsterdam, The Netherlands) for the shared care of the two Dutch patients with MVID, and the Dutch Digestive Diseases Foundation (MLDS) for supporting the national collaboration for patients with intestinal failure. We thank Julius Baller and Mathieu Platteel for expert technical assistance, and Hilda Keuning for analysis of histological findings. Control DNA samples were provided by Yvonne Vos (Dutch and Moroccan) and Anna Rybak (Polish). Hanna Romanowska collected DNA samples from the Polish patient and her family. We are indebted to Henkjan Verkade for critically reading the manuscript. This research was sponsored by the Dutch Digestive Foundation (MLDS), the Jan Kornelis de Cock Foundation, and De Drie Lichten Foundation. MG was supported by the Ubbo Emmius Foundation. CW was supported by a grant from the Netherlands Organization for Scientific Research (NWO-VICI). SvIJ and ER were supported by the Royal Dutch Academy of Arts and Sciences.

### References

1. Phillips A. D., Jenkins P., Raafat F., Walker-Smith J. A. (1985) Congenital microvillous atrophy: specific diagnostic features. *Arch. Dis. Child*, **60**, 135-140.
2. Phillips A. D., Schmitz J. (1992) Familial microvillous atrophy: a clinicopathological survey of 23 cases. *J. Pediatr. Gastroenterol. Nutr.*, **14**, 380-396.
3. Cutz E., Rhoads J. M., Drumm B., Sherman P. M., Durie P. R., Forstner G. G. (1989) Microvillus inclusion disease: an inherited defect of brush-border assembly and differentiation. *N. Engl. J. Med.*, **320**, 646-651.

4. Sherman P. M., Mitchell D. J., Cutz E. (2004) Neonatal enteropathies: defining the causes of protracted diarrhea of infancy. *J. Pediatr. Gastroenterol. Nutr.*, **38**, 16-26.
5. Goulet O., Ruummele F., Lacaille F., Colomb V. (2004) Irreversible intestinal failure. *J. Pediatr. Gastroenterol. Nutr.*, **38**, 250-269.
6. Ruummele F. M., Schmitz J., Goulet O. (2006) Microvillous inclusion disease (microvillous atrophy). *Orphanet. J. Rare. Dis.*, **1**, 22.
7. Iancu T. C., Mahajnah M., Manov I., Shaoul R. (2007) Microvillous inclusion disease: ultrastructural variability. *Ultrastruct. Pathol.*, **31**, 173-188.
8. Ameen N. A., Salas P. J. (2000) Microvillus inclusion disease: a genetic defect affecting apical membrane protein traffic in intestinal epithelium. *Traffic*, **1**, 76-83.
9. Michail S., Collins J. F., Xu H., Kaufman S., Vanderhoof J., Ghishan F. K. (1998) Abnormal expression of brush-border membrane transporters in the duodenal mucosa of two patients with microvillus inclusion disease. *J. Pediatr. Gastroenterol. Nutr.*, **27**, 536-542.
10. Muller T., Hess M. W., Schiefermeier N., Pfaller K., Ebner H. L., Heinz-Erian P., Ponstingl H., Partsch J., Rollingshoff B., Kohler H., *et al.* (2008) MYO5B mutations cause microvillus inclusion disease and disrupt epithelial cell polarity. *Nat. Genet.*, **40**, 1163-1165.
11. Erickson R. P., Larson-Thome K., Valenzuela R. K., Whitaker S. E., Shub M. D. (2008) Navajo microvillous inclusion disease is due to a mutation in MYO5B. *Am. J. Med. Genet. A*, **146A**, 3117-3119.
12. Ruummele F. M., Muller T., Schiefermeier N., Ebner H. L., Lechner S., Pfaller K., Thoni C. E., Goulet O., Lacaille F., Schmitz J., *et al.* (2010) Loss-of-function of MYO5B is the main cause of microvillus inclusion disease: 15 novel mutations and a CaCo-2 RNAi cell model. *Hum. Mutat.*, **31**, 544-551.
13. Prekeris R., Klumperman J., Scheller R. H. (2000) A Rab11/Rip11 protein complex regulates apical membrane trafficking via recycling endosomes. *Mol. Cell*, **6**, 1437-1448.
14. Holmes K. C., Schroder R. R., Sweeney H. L., Houdusse A. (2004) The structure of the rigor complex and its implications for the power stroke. *Philos. Trans. R. Soc. Lond B Biol. Sci.*, **359**, 1819-1828.
15. Coureux P. D., Sweeney H. L., Houdusse A. (2004) Three myosin V structures delineate essential features of chemo-mechanical transduction. *EMBO J.*, **23**, 4527-4537.
16. Tang S., Liao J. C., Dunn A. R., Altman R. B., Spudich J. A., Schmidt J. P. (2007) Predicting allosteric communication in myosin via a pathway of conserved residues. *J. Mol. Biol.*, **373**, 1361-1373.
17. Cecchini M., Houdusse A., Karplus M. (2008) Allosteric communication in myosin V: from small conformational changes to large directed movements. *PLoS. Comput. Biol.*, **4**, e1000129.
18. Isken O., Maquat L. E. (2008) The multiple lives of NMD factors: balancing roles in gene and genome regulation. *Nat. Rev. Genet.*, **9**, 699-712.
19. Lapierre L. A., Kumar R., Hales C. M., Navarre J., Bhartur S. G., Burnette J. O., Provance D. W., Jr., Mercer J. A., Bahler M., Goldenring J. R. (2001) Myosin vb is associated with plasma membrane recycling systems. *Mol. Biol. Cell*, **12**, 1843-1857.
20. Swiatecka-Urban A., Talebian L., Kanno E., Moreau-Marquis S., Coutermarsh B., Hansen K., Karlson K. H., Barnaby R., Cheney R. E., Langford G. M., *et al.* (2007) Myosin Vb is required for trafficking of the cystic fibrosis transmembrane conductance regulator in Rab11a-specific apical recycling endosomes in polarized human airway epithelial cells. *J. Biol. Chem.*, **282**, 23725-23736.
21. Nedvetsky P. I., Stefan E., Frische S., Santamaria K., Wiesner B., Valenti G., Hammer J. A., III, Nielsen S., Goldenring J. R., Rosenthal W., Klussmann E. (2007) A Role of myosin Vb and Rab11-FIP2 in the aquaporin-2 shuttle. *Traffic*, **8**, 110-123.
22. Roland J. T., Kenworthy A. K., Peranen J., Caplan S., Goldenring J. R. (2007) Myosin Vb interacts with Rab8a on a tubular network containing EHD1 and EHD3. *Mol. Biol. Cell*, **18**, 2828-2837.
23. Hoekstra D., Tyteca D., van IJendoorn S. C. (2004) The subapical compartment: a traffic center in membrane polarity development. *J. Cell Sci.*, **117**, 2183-2192.
24. van IJendoorn S. C. (2006) Recycling endosomes. *J. Cell Sci.*, **119**, 1679-1681.

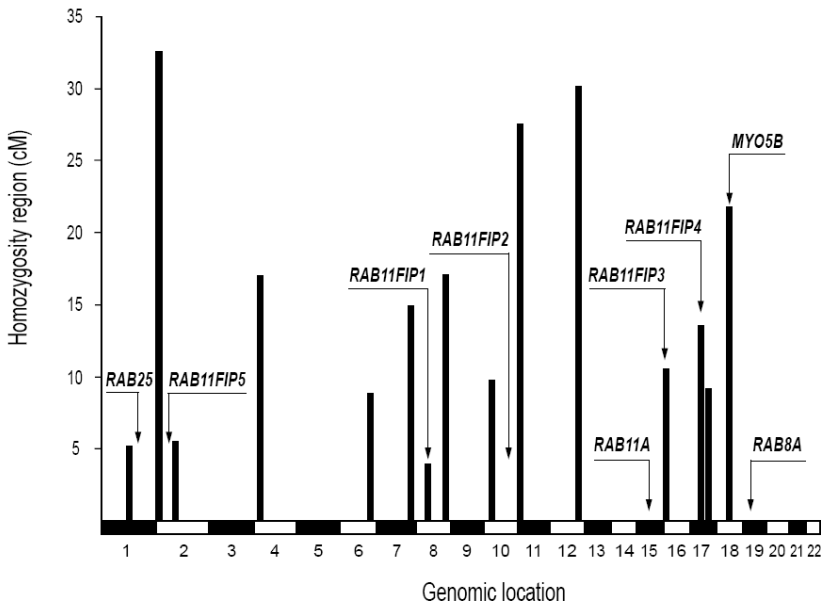
## Unpublished Data

### Homozygosity mapping

Genome-wide homozygosity mapping was performed using 200ng of genomic DNA isolated from peripheral blood on Infinium HumanLinkage-12 BeadChip (Illumina, USA) according to the manufacturer's protocol. The DNA samples were hybridized to a panel of 6,090 SNP markers on the Illumina HumanLinkage-12 BeadChip.

### Results

Homozygosity mapping was performed using the DNA of patient 9, who is the son of first-degree cousins. Homozygous regions with considerable genetic distances and physical sizes were located (this text : Figure 1). Comparison of these positional candidate regions with the genomic locations of known apical recycling endosome-associated and functional candidate proteins yielded *MYO5B* as a positional-functional candidate gene in this patient (this text : Figure 1). Sequencing of the *MYO5B* gene in patient 9 revealed a homozygous stop codon in exon 33 (p.Q1456X; c.4366C>T) (main text : Figure 1A).



**Figure 1. Homozygosity mapping in patient # 9 .** The position and length of the homozygous segments along the genome are indicated together with the location of candidate genes from the apical recycling endosome pathway.



*True autosomal dominant inheritance of FMF caused by a mutation in exon 8 of the MEFV gene*

*Monique Stoffels<sup>1,2,3</sup>, Agata Szperl<sup>4</sup>, Marielle van Gijn<sup>5</sup>, Mihai G. Netea<sup>1,2,3</sup>, Theo S. Plantinga<sup>1,3</sup>, Marcel van Deuren<sup>1,3</sup>, Sylvia Kamphuis<sup>7</sup>, Helen Lachmann<sup>8</sup>, Edwin P.J.G. Cuppen<sup>5</sup>, Wigard P. Kloosterman<sup>5</sup>, Joost Frenkel<sup>6</sup>, Cleo C. van Diemen<sup>4</sup>, Cisca Wijmenga<sup>4</sup>, Anna Simon<sup>1,2,3,\*</sup>, and Jos W.M. van der Meer<sup>1,2,3</sup>*

*<sup>1</sup> Department of General Internal Medicine, Radboud University Nijmegen Medical Centre, Nijmegen, the Netherlands*

*<sup>2</sup> Nijmegen Centre for Molecular Life Sciences, Radboud University Nijmegen Medical Centre, Nijmegen, the Netherlands*

*<sup>3</sup> Nijmegen Centre for Infection, Inflammation and Immunity (N4i), Nijmegen, the Netherlands*

*<sup>4</sup> Department of Genetics, University of Groningen and UMC Groningen, Groningen, the Netherlands*

*<sup>5</sup> Department of Medical Genetics, UMC Utrecht, Utrecht, the Netherlands*

*<sup>6</sup> Division of Pediatrics, UMC Utrecht, Utrecht, the Netherlands*

*<sup>7</sup> Sophia Children's Hospital, Erasmus UMC, Rotterdam, the Netherlands*

*<sup>8</sup> University College London Medical School, Royal Free Campus, National Amyloidosis Centre; Centre for Amyloidosis and Acute Phase Proteins, Division of Medicine, London, UK*

*This manuscript has been submitted*

## Abstract

Familial Mediterranean Fever (FMF) is an autosomal recessive auto-inflammatory disease, usually caused by mutations in exon 10 of the *MEFV* gene, coding for pyrin. Standard genetic testing could not provide a satisfactory diagnosis in a female patient with an autosomally dominant inherited periodic fever syndrome. She came from a non-consanguineous family of British descent. Whole exome sequencing revealed a novel missense sequence variant not seen in around 6,800 controls, mapping to exon 8 of the *MEFV* gene (c.1730C>A; p.T577N). It co-segregated perfectly with the disease in this family. Other mutations at the same amino acid (c.1730C>G; p.T577S; c.1729A>T; p.T577S) were found in a family of Turkish descent, with autosomal dominant inheritance of an FMF-like phenotype, and in a Dutch patient. A mutation (c.1729A>G; p.T577A) was also detected in two Dutch siblings suffering from episodes of inflammation of varying severity but not resembling FMF. Peripheral-blood mononuclear cells (PBMCs) from one patient of our index family revealed increased basal IL-1 $\beta$  mRNA levels and cytokine responses after lipopolysaccharide (LPS) stimulation. Responses normalized under colchicine treatment. We conclude that heterozygous mutations at amino acid position 577 of pyrin can induce an autosomally dominant auto-inflammatory syndrome. This suggests that T577, located between the coiled coil- and the C-terminal B30.2/SPRY domains, is of crucial importance for pyrin function.

## Introduction

Auto-inflammatory disorders are disorders of the innate immune system characterized by recurrent episodes of fever and systemic inflammation, with fever and localized inflammation predominantly affecting serosal surfaces, skin, joints and eyes. Attacks are often self-resolving and there are no signs of auto-antibodies or infection (1, 2).

Auto-inflammatory syndromes may be *inherited*, such as Familial Mediterranean Fever (FMF), TNF receptor–associated periodic syndrome (TRAPS), mevalonate kinase deficiency (MKD), cryopyrin-associated periodic syndromes (CAPS), Blau syndrome, deficiency of the interleukin-1 receptor antagonist (DIRA), pyogenic arthritis, pyoderma gangrenosum, and acne (PAPA), or *acquired*, such as periodic fever with aphthous stomatitis, pharyngitis, and cervical adenitis (PFAPA), and Schnitzler syndrome. There are some similarities in symptoms, but the clinical picture, the mode of inheritance, duration and frequency of attacks are often different (2, 3). In addition, many patients suspected of having an auto-inflammatory disease, do not fulfill the diagnostic criteria for the known auto-inflammatory syndromes and/or test negative for the known mutations in genes associated with inherited auto-inflammatory syndromes. Correctly diagnosing these patients can therefore be challenging, as shown by the long period often needed for diagnosis (4).

In a family with an autosomally dominant auto-inflammatory disorder, standard genetic testing did not provide a satisfactory diagnosis. We aimed to reveal the genetic cause of the autosomally dominant inflammatory symptoms in this family by using more advanced genetic methodology, such as a hypothesis-free, exome-wide sequencing approach.

## Material and Methods

### Patients

*Family 1.* A female patient (IV: 3) from a non-consanguineous family of British descent with an apparently autosomally dominant inherited periodic fever syndrome (family 4, Figure 1A) suffered from recurrent episodes of synovitis, pleuritis or peritonitis and skin rash, from the age of 6 years. Her grandfather (II: 1) suffered from inflammatory episodes of fever, arthritis, pleuritis, peritonitis, and skin rash from the age of 10. His father, two of his siblings, and his daughter



(mother of the proband; III: 2) had similar attacks. They experienced beneficial responses to colchicine. However, initial DNA mutation screening in at least 2 family members for *MEFV* exons 2, 5 and 10, as well as *TNFRSF1A* exons 2, 3, 4, 5, 6 and 7, and *NLRP3* did not reveal any mutations.

We characterized the immune responses of peripheral-blood mononuclear cells (PBMCs) from one affected family member (III: 2), and one healthy control subject. We carried out genetic analyses of one affected and one unaffected family members.

To check if the variant was a common polymorphism, we investigated 396 healthy unrelated Dutch controls (from the genetic diagnostic department, Utrecht), 500 unrelated Dutch controls from the GoNL (Genome of the Netherlands) project, from the 1000 Genomes Project (120 CEU and 118 YRI controls ) (5), 200 Danish controls (6), 120 UK controls and an NHLBI GO Exome Sequencing Project (ESP) dataset of 5,400 individual controls.

#### **Exome sequencing: library generation, reference alignment and variant calling**

The genomic DNA samples were randomly fragmented using nebulization. Barcoded adapters were ligated to both ends of the resulting fragments, according to the standard NEBNext® DNA Library Prep Master Mix Set for Illumina® protocol (New England Biolabs, UK) (7). Fragments with an average insert size of 220 bp were excised using the Caliper XT gel system and the extracted DNA was amplified by polymerase chain reaction (PCR). The quality of the product was verified on the BioRad Experion instrument. If the quality of the product met the criteria, the product was multiplexed in an equimolar pool of four similar products. This pool was hybridized to the Agilent SureSelect All exon V2 kit, according to the manufacturer's protocol. After amplification of the enriched products with PCR, the quality of the products was verified on the BioRad Experion instrument followed by paired-end sequencing on the HiSeq2000 with 100 bp reads. Image files were processed using standard Illumina base-calling software and the generated reads were ready for downstream processing after demultiplexing.

Reads were aligned to the human reference genome, build 37, using Burrows-Wheeler Alignment (BWA) (8). To clean the aligned data and perform variant calling, we applied Picard duplicate removal and the Genome Analysis Toolkit (GATK) (9) quality score recalibration, indel realignment, and unified genotyper. Using snpEff

(<http://snpeff.sourceforge.net>) the variants were annotated with information from dbSNP132, the 1000 Genomes Project (phases 1, 2 and 3) and Ensembl, build 37.64. These programs are all integrated into the MOLGENIS Compute pipeline developed by the Genomics Coordination Center, Department of Genetics, UMCG and University of Groningen, the Netherlands (10) (<http://wiki.gcc.rug.nl/wiki/GccStart>). Pathogeneity predictions for variants were obtained from PolyPhen 2.0, SIFT, and Align GVD (11 - 14).

### **Exome sequencing: step-wise filtering of the sequence data**

For our analysis we have used a “linkage-based strategy” analysis (15). We have chosen all sequence variants (SVs) shared between two affected, related individuals (21,905 SVs). We excluded all variants that: (1) were reported in the 131dbSNP (including 1000 Genomes Project) and in private sets of two unrelated samples (2,225 SVs left), (2) mapped to the intronic regions of the genes, except to splice sites (565 SVs left), (3) were present in the olfactory receptor genes (16) and that had a high copy number genes (17). We further reduced our dataset to 151 SVs by removing all variants in the homozygous stage (dominant model of disease), present on the X chromosome (autosomal model), and by using an updated SeattleSeq Annotation tool (18), removing additional variants present in the 1000 Genomes Project but not yet present in 131 dbSNP database, at this point of analysis. Finally, we removed all the variants predicted to be “benign” by PolyPhen implemented in SeattleSeq Annotation tool and this left us with 125 SVs for further consideration.

### **Sequencing of 120 inflammasome related genes**

We generated a barcoded, whole genome, fragment library for the patient in family 4. The barcoded library was enriched from the coding regions of 120 inflammasome genes using a custom Agilent 1M microarray and the enriched library was sequenced on the SOLiD4 sequencer, as described previously (19).

### **Validation of mutations**

Variations were validated by direct Sanger sequencing and analyzed using the DNA Variant Analysis software (Mutation Surveyor). To validate the presence of mutations in *MEFV* in affected patients of family 1, conventional PCR and Sanger

sequencing of exon 8 were performed. DNA was isolated from whole blood using standard procedures. Primer sequences are available upon request.

### ***In vitro* cytokine production**

PBMCs were isolated using Ficoll-Paque Plus (Bio-Sciences AB).  $5 \times 10^5$  cells in RPMI 1640 medium (Dutch modification; Invitrogen, Paisley, Scotland), supplemented with 1 mM sodium pyruvate (Sigma-Aldrich), 2 mM L-Glutamine (Merck) and 50  $\mu\text{g}/\text{ml}$  gentamicin (Centrafarm), were incubated in round-bottomed 96-well plates (Greiner) at 37°C. After 24 hours of incubation with various stimuli, supernatants were collected and stored at -80°C until further analysis. In another experimental set-up, cells were stimulated for 3 hours with 1  $\mu\text{g}/\text{ml}$  lipopolysaccharide (LPS) purchased from Sigma (Escherichia coli serotype 055:B5) and purified as described elsewhere (20), followed by 15 min 1 mM ATP stimulation to induce IL-1 $\beta$  release (21).

### **Cytokine assays**

Cytokine concentrations in supernatants were measured using commercial enzyme-linked immunosorbent assay (ELISA) kits from R&D Systems (IL-1 $\alpha$ , IL-1 $\beta$ ) or PeliPair Reagent sets from Sanquin (IL-6), according to manufacturer's instructions.

### **qRT-PCR**

One million freshly isolated PBMCs were incubated with various stimuli. After incubation at 37°C, total RNA was extracted by using TRIzol reagent (Invitrogen). Isolated RNA was subjected to DNase treatment (Ambion® DNA-free™ Kit, Invitrogen) and 0.5  $\mu\text{g}$  RNA was reverse-transcribed into cDNA using iScript cDNA Synthesis Kit (Bio-Rad, Hercules, California). qRT-PCR for human IL-1 $\beta$  and  $\beta$ 2-microglobulin was performed using an Applied Biosystems 7300 real-time PCR system. Primer sequences are available upon request.

### **Immunoblotting for Pypin**

For immunoblotting,  $5 \times 10^6$  freshly isolated PBMCs were lysed in 100  $\mu\text{l}$  of lysis buffer (50 mM Tris (pH 7.4), 150 mM NaCl, 2 mM EDTA, 2 mM EGTA, 10% glycerol, 1% Triton X-100, 40 mM  $\beta$ -glycerophosphate, 50 mM sodium fluoride, 200  $\mu\text{M}$  sodium vanadate), supplemented with protease inhibitor cocktail (Roche)

and the homogenate was frozen. Upon further processing, the supernatant was used for Western blot analysis. Equal amounts of protein were subjected to SDS-PAGE, after which proteins were transferred to a nitrocellulose membrane. Membranes were blocked in 5% (wt/vol) non-fat dried milk (NFDM) in Tris-buffered saline, containing 0.1% Tween-20 (TBS-T), followed by overnight incubation at 4°C with pyrin antibody (sc-30421; Santa Cruz Biotechnology) in 5% NFDM in TBS-T. Membranes were washed with TBS-T and then incubated with HRP-conjugated rabbit anti-goat antibody (DAKO) in 5% NFDM in TBS-T for 1 hour at room temperature. Blots were developed with enhanced chemiluminescence (ECL, GE Healthcare) according to the manufacturer's instructions.

## Results

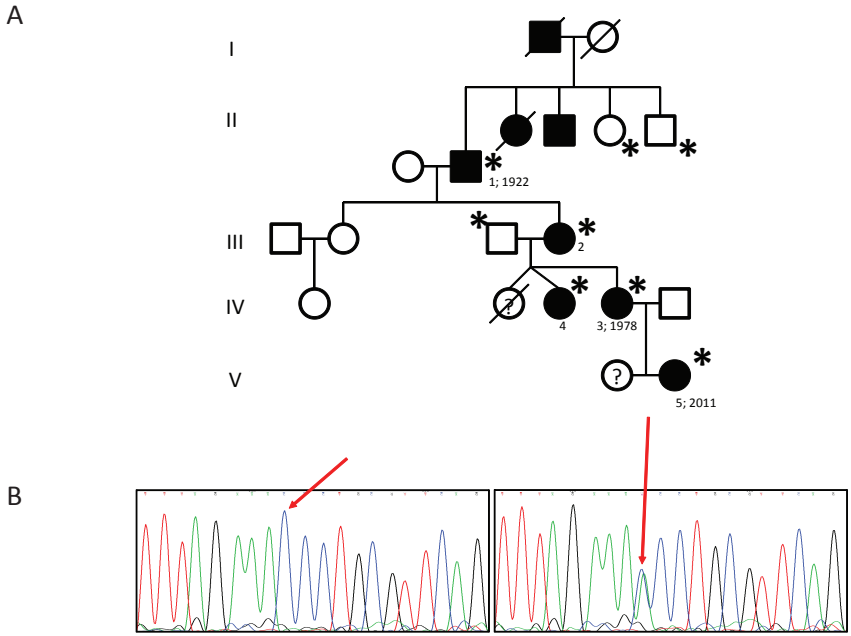
### Genetic cause of the auto-inflammatory disease in family 1

To search for gene candidates, we employed whole exome sequencing in two patients from family 1 (II: 1 and IV: 3). The raw sequencing was aligned to the human reference build 37, providing 27 Gb of high quality reads. The mean coverage on target was 76- and 135-fold for patients II: 1 and IV: 3, respectively. 92% and 93%, respectively, of the targeted region was covered at least 10 times, indicating high quality data. The concordance of sequenced data with genotyping data (HumanCytoSNP-12 BeadChip, Illumina) for the same individuals was 99.9%. On average, we have identified 31,600 sequencing variants (SVs) of high quality per individual, which were then used for the step-wise filtering procedure.

Filtering out all variants present in 500 unrelated Dutch controls from the GoNL project and the complete CEU and YRI population from the 1000 Genomes Project resulted in a list of 125 gene candidates with mutations present in the two patients. One of the candidates was a mutation in exon 8 of the *MEFV* gene, which encodes the pyrin protein, c.1730C>A, resulting in a missense mutation, p.T577N, in the pyrin protein.

Considering the patients' good response to colchicine, the mutation in the *MEFV* gene was the most likely candidate. We further tested family members for how this variant segregated with disease. As shown in figures 1A and B, our sequencing confirmed the co-segregation of the mutation with the affected family members. More importantly, this mutation was not found in 396 healthy Dutch controls, 200 Danish controls, 120 UK controls, or in the ESP dataset, nor in the

500 Dutch controls from the GoNL project. Neither could we detect this mutation in over 1,000 patients with unexplained auto-inflammatory symptoms known to the Department of Medical Genetics, UMC Utrecht or the Department of General Internal Medicine, Nijmegen.



**Figure 1. Mutation analysis in family 1.** (A) Pedigree of family 1. Affected family members (according to symptoms described) are indicated in black; males and females are indicated by squares and circles, respectively. \* indicates that clinical data were confirmed with sequencing data. Question marks indicate that no information was available on these individuals. (B) The mutation found in patients in family 1 is a heterozygous autosomal dominant c.1730C>A mutation, resulting in a p.T577N amino acid change at a position in front of the B30.2 domain of the pyrin protein. Since the crystal structure has not yet been fully resolved, we could not apply any modeling to study the possible consequences of this mutation.

**Pyrin mutations in the same amino acid position in more families with auto-inflammatory diseases**

At the Department of Medical Genetics, UMC Utrecht, three different mutations, never described before, affecting the same amino acid location of pyrin had previously been detected in three other patients or families.

A c.1730C>G (p.T577S) mutation in *MEFV* (without other *MEFV* gene mutations) was found in a family of Turkish descent, who also presented with autosomally dominant inheritance of FMF-like phenotype (figure 2A). Moreover, a c.1729A>T (p.T577S) mutation in association with another non-classic *MEFV* gene

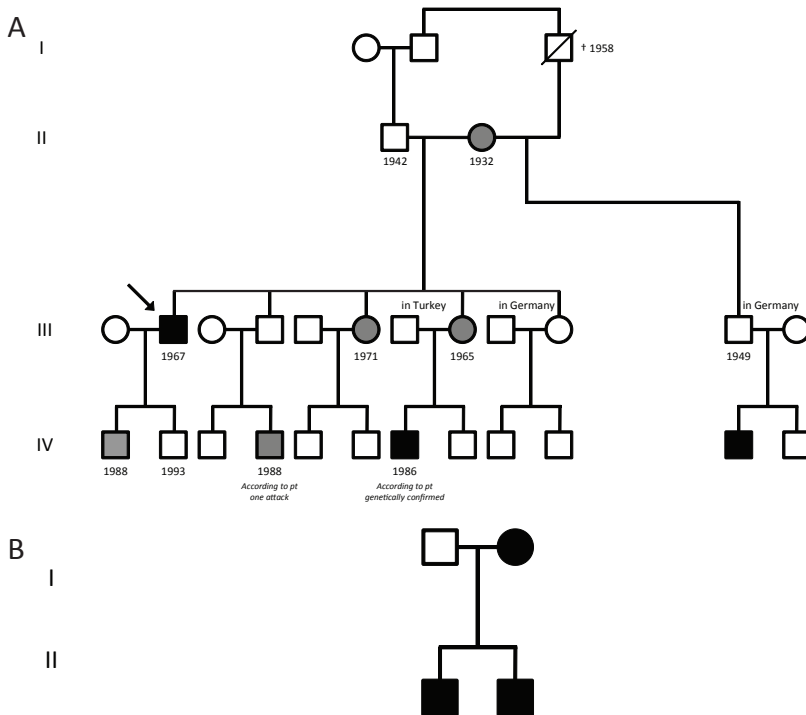


Figure 2. Pedigrees of (A) family 2 and (B) family 4

variant (c.800C>T (p.T267I)) was detected in a Dutch patient, to confirm the clinical diagnosis of FMF.

Finally, sequencing of 120 inflammasome-related genes in a Dutch family with a very special clinical presentation also revealed a c.1729A>G (p.T577A) mutation, in association with another non-classic *MEFV* gene variant (369/408 complex allele). No other predicted pathogenic variants were detected (figure 2B). Since the age of one, the index case in this particular family has suffered from skin abnormalities, periodic painful swollen joints, occasional fever, growth- and developmental delay, a substantially enlarged spleen, a mildly enlarged liver, and extreme anemia (Hb 3.5 mmol/l). Standard genetic testing for CAPS and DIRA was negative. Under anakinra treatment (a synthetic interleukin-1 receptor antagonist blocking the effects of human interleukin-1) this patient is now doing well (Hb 6.5 mmol/l within one month) and is catching up on her psychomotor delay. Her mother also has a periodic inflammatory presentation, which started at age 6 years and with less severe anemia. She is considerably feeling better on anakinra treatment. The other son was put on colchicine treatment for 4-6 weeks

within one year of birth, but this was not successful, and anakinra treatment was started. This treatment proved successful and is now being optimized.

### Functional studies

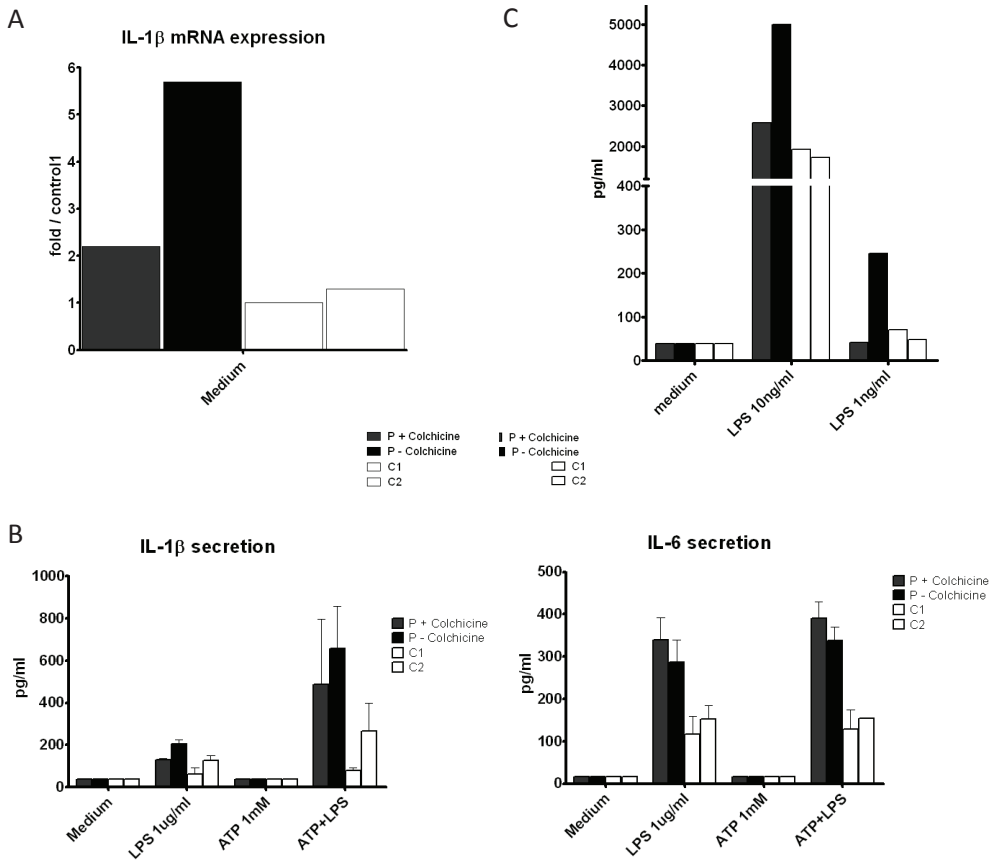
Since the T577 mutations are located in exon 8, just before the PRY-SPRY domain of pyrin, we hypothesized that it might alter the structure of the protein. However, a Western blot on PBMC lysates from patient III:2 from family 1 showed no differences in banding pattern compared to the control (not shown), indicating that the mutation did not affect the epitopes recognized by the pyrin antibody.

To study the functional consequences of the T577N mutation, we isolated PBMCs from patient III: 2 (family 1) and a healthy volunteer. First, we examined the IL-1 $\beta$  mRNA levels. As shown in figure 3A, PBMCs obtained from the T577N patient without colchicine treatment showed an almost six times higher expression of basal IL-1 $\beta$  mRNA. While the patient was being treated, the expression levels dropped and resembled control levels, indicating that colchicine treatment was acting on the expression of IL-1 $\beta$  mRNA and that this mutation caused higher IL-1 $\beta$  expression levels.

Next, to evaluate the functional consequences of T577N at the protein level, LPS stimulation studies were performed. As shown in figure 3B, the patient's cells secreted more of the pro-inflammatory cytokines, IL-1 $\beta$  and IL-6, after 3 hours of LPS and LPS+ATP stimulation than controls, even under colchicine treatment. After 24 hours incubation in a medium, and with low dose LPS (figure 3C), T577N PBMCs obtained from the patient without treatment showed a higher production (not shown) and secretion of cytokines (IL-1 $\beta$ , IL-1 $\alpha$ , IL-6), compared to the control. Treatment with colchicine normalized the secretion of IL-1 $\beta$  to control levels. In response to low dose LPS, only the PBMCs of the patient not taking colchicine were able to secrete excessive amounts of cytokines, illustrating that the cells bearing the mutation were more easily triggered to produce cytokines.

### Discussion

We present a family with a truly autosomal dominant, auto-inflammatory disorder, caused by a single c.1730C>A mutation in exon 8 of the *MEFV* gene that results in a missense mutation (p.T577N) in the pyrin protein. The T577N mutation seems to be severe enough to cause an auto-inflammatory phenotype without the presence



### Figure 3. Ex vivo inflammasome stimulation

PBMCs were incubated for 3 hours with 1  $\mu$ g/ml LPS or RPMI, after which the medium was removed. Cells were incubated for an extra 15 minutes with either RPMI or 1 mM ATP (A, B). In another experiment (C), cells were stimulated for 24 hours with 1 or 10 ng/ml lipopolysaccharide (LPS). After incubation, the plates were spun, the supernatants collected and mRNA was isolated. IL-1 $\beta$  mRNA expression of one patient from family 1, under colchicine treatment (grey bar) or not (black bar), and a control person (white bars). Basal mRNA levels for IL-1 $\beta$  in the patient without treatment are 5.7 times higher than control basal IL-1 $\beta$  mRNA levels. Under colchicine treatment, these levels seemed to normalize to control values. (B) There was a trend to higher IL-1 $\beta$  secretion in the patient, but treatment apparently played no role here. (C) The untreated patient was more sensitive to LPS stimulation, which was even more apparent at low dose LPS. IL-1 $\beta$  secretion normalized to control levels when the patient was treated with colchicine.

of any other mutations in this gene. We identified three more families who also harbored a T577 mutation. The fact that four rare, novel, DNA variants, affecting the same amino acid, were detected in auto-inflammatory patients from different populations, suggests that T577 is of crucial importance in pyrin function. The classic disease caused by pyrin mutations is Familial Mediterranean Fever, which is inherited as an autosomally recessive disease and is due to two mutations in homozygous or compound heterozygous form, most often located in exon 10.



To date, the precise structure and function of the pyrin protein have not been fully resolved. The pyrin protein consists of an N-terminal Pyrin domain, a B-box, bZIP basic, coiled-coil domains and a B30.2/SPRY domain (figure 1C) (22-24). To investigate the effects of the mutations found in FMF, Weinert et al. 2009 (25) resolved the crystal structure of the C-terminal B30.2 domain at 1.35-Å resolution, where about two-thirds of the mutations are located (exon 10), leading to the conclusion that many mutations are close to a predicted peptide-binding site and suggesting that they are likely involved in altering the binding properties of the B30.2 domain. T577 is located between the coiled-coiled domain and the B30.2 domain and structural modeling based on an experimentally determined crystal structure could suggest the implications of a mutation at this site in the pyrin protein. We attempted to model the mutated protein, but this proved impossible because the pyrin crystal structure is not completely resolved. The T577N mutation most probably leads to a structural change in the pyrin protein that either affects its interaction with other proteins, or impairs its function, such that it cannot be compensated for by the correct protein product encoded by the other allele. The fact that one of the families presented with an auto-inflammatory disease with symptoms that are atypical for FMF also indicates that T577 is important for pyrin function, and might result in other presentations of auto-inflammation.

There is still no agreement on how pyrin dysfunction contributes to auto-inflammation. At least part of the protein was found in the nucleus of granulocytes and dendritic cells (26, 27), where it might activate NF-κB. Pyrin was also found in the cytoplasm of monocytes, interacting with tubulin and co-localizing with microtubules (28). Through the pyrin domain, it can interact with ASC (apoptosis-associated speck-like) protein, which is involved in inflammasome complex formation (29), thereby activating IL-1β. The exact role of pyrin in IL-1β activation is not clear, because experimental findings fit in with two possible mechanisms: either pyrin inhibits IL-1β activation by competing with caspase-1 for ASC (30, 31), or it forms an inflammasome complex by itself (32). There are groups of studies supporting both these contradictory ideas of the functional inhibition (30, 31, 33-35) or activation (32, 36) of IL-1β production and release; these studies are reviewed more extensively by Masters et al (2).

Our cellular studies show that colchicine treatment *in vivo* inhibits the expression level of IL-1β (by almost six-fold compared to a control without

treatment), although further research is needed to pinpoint the mechanism through which colchicine exerts its effect. Our data also show that the T577N mutation results in increased cytokine secretion after LPS stimulation and that this normalizes when the patient is on colchicine treatment. However, from these data, we cannot conclude whether the increased cytokine production is due to a gain-of-function, or to a loss of the inhibitory effect of pyrin. Whatever the mechanism, functionally the T577 mutation has a clear pro-inflammatory effect.

In the era before the discovery of the *MEFV* gene mutations, autosomally dominant FMF was described (37). Because genetic confirmation was not possible at that time, the results have to be interpreted with caution. Nevertheless, after the discovery of pyrin in 1997 (38), reports about autosomally dominant FMF have persisted. There are a couple of possible explanations: the simplest one is that the second *MEFV* mutation was not found. When only the known, most common mutations are screened for, other rare or novel variants will not be detected, and differences in genomic re-arrangements such as large deletions or copy number variations (CNV) will also not be found. However, according to Van Gijn et al, single or multi-exon CNVs of the *MEFV* gene do not contribute to the mutation spectrum (39). Another possibility is that the second mutation is located in the intronic or promoter regions of *MEFV*, which are not routinely screened for in FMF patients. In addition, the presence of a mutation in other auto-inflammatory genes (40-43), or modifications at epigenetic (44), post-translational level have been reported to occur in combination with one *MEFV* mutation. In this study, we employed whole exome sequencing, largely so that we would avoid missing mutations in the coding part of the *MEFV* gene, or other auto-inflammatory genes. However, it is still possible that we missed another mutation due to the design of the enrichment or to some problematic regions in the genome that were not well covered (15).

Although this screening does not take into account the intronic and promoter regions, it is very unlikely that a second intronic mutation would be perfectly co-segregating with the disease, in addition to the T577 mutation.

Another phenomenon that might explain why only one mutation has been found in some patients, is pseudo-dominance, which might arise in the offspring of a heterozygous carrier and a homozygous patient. Pseudo-dominant inheritance can occur in populations with a high carrier frequency and consanguinity, which both apply to the Mediterranean populations (45). This appears to be the case

in several studies (46-48). In our study, two of the four families are of Dutch and British ancestry, which is atypical for FMF and it therefore seems highly unlikely there would be a second mutation in successive generations. Furthermore, after applying our filtering criteria to the exome sequencing data, we did not find any other candidate mutations in *MEFV* in two patients.

A different explanation for not finding a second mutation could be the phenomenon of mosaicism (49-51), whereby individuals carry cells of different genotypes. For instance, it has been reported that *NLRP3/CIAS1* mosaicism is a significant cause of cryopyrin-associated periodic syndromes (CAPS). However, this has not been reported for FMF, and mosaicism cannot explain the inheritance seen in our family.

As a final reason for not finding a second mutation, we would like to propose that FMF (or “pyrin-associated periodic syndrome”) can be inherited either recessively or dominantly, depending on the type of mutation. There are a few recent studies that used extensive screening methods and reported a dominant inheritance. Aldea et al reported a heterozygous H478Y, without any other affected gene (52). However, according to the authors, as the phenotype was not typically FMF, this mutation might cause a so-far-undescribed FMF-like phenotype. In three other recent studies, families with autosomal dominant inheritance were reported (40, 46, 47) in which dominant inheritance linked to the deletion of M694 in exon 10 or a combination of mutations (with or without M694) within one allele, also known as a complex allele, was found. M694 mutations are known to cause a more severe phenotype. The R653H mutation was reported but not in combination with another mutation, however it is located in exon 10. Our study shows a mutation in a previously undescribed region with regard to mutations, which is apparently also strong enough for only one mutated allele to cause disease, as also reported for the delM694.

It is of great importance that FMF patients are diagnosed correctly so that they can be treated appropriately. This is important in FMF to prevent the development of amyloidosis (1). For example, subject III:2 of family 1 was about 70 years old before we were able to pinpoint the true cause of her auto-inflammatory symptoms. In the meantime she had been incorrectly diagnosed several times, because symptoms of FMF overlap with other diseases, such as rheumatoid arthritis. As a result, she also suffered side-effects from an inappropriate treatment.

In summary, we have identified a family with a truly autosomal dominant inheritance of FMF, caused by a single amino acid transition, on a single allele. This T577 mutation is not located in exon 10, and is linked to autosomal dominant inheritance of FMF, even in non-classical populations. Our finding in this family is corroborated by finding mutations at position 577 in three other families with autosomally dominant auto-inflammation. Our results offer support for the use of exome screening in non-Mediterranean families with unexplained dominant auto-inflammatory diseases, who have a good response to colchicine. Exome screening can lead to a correct diagnosis and appropriate treatment. Apparently, T577 plays an important role in pyrin function, since only one affected allele is enough to cause the FMF phenotype, and it cannot be rescued by the product from the healthy allele.

### **Acknowledgements**

Dr. Simon was supported by an NWO-VIDI grant (Netherlands Organization for Scientific Research) and Dr. van Gijn by a Horizon grant from the Netherlands Genomics Initiative and ZonMw. We thank our patients and their families; Dr. van de Vosse and colleagues (LUMC) for making their lab available, and Prof. Vriend (CMBI) for helping with the modeling. We also thank Ivo Renkens (UMC Utrecht, SOLiD sequencing), Les Nijman (UMC Utrecht, bioinformatics analysis of inflammasome screen), and José van de Belt, library preparation for the inflammasome screen). We thank Jackie Senior, UMCG, for editing the manuscript.

## Reference List

1. Lachmann H. J. (2011) Clinical immunology review series: An approach to the patient with a periodic fever syndrome. *Clin. Exp. Immunol.*, **165**, 301-309.
2. Masters S. L., Simon A., Aksentjevich I., Kastner D. L. (2009) Horror autoinflammaticus: the molecular pathophysiology of autoinflammatory disease (\*). *Annu. Rev. Immunol.*, **27**, 621-668.
3. Goldbach-Mansky R. (2012) Immunology in clinic review series; focus on autoinflammatory diseases: update on monogenic autoinflammatory diseases: the role of interleukin (IL)-1 and an emerging role for cytokines beyond IL-1. *Clin. Exp. Immunol.*, **167**, 391-404.
4. van der Hilst J. C., Bodar E. J., Barron K. S., Frenkel J., Drenth J. P., van der Meer J. W., Simon A. (2008) Long-term follow-up, clinical features, and quality of life in a series of 103 patients with hyperimmunoglobulinemia D syndrome. *Medicine (Baltimore)*, **87**, 301-310.
5. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061-1073.
6. Li Y., Vinckenbosch N., Tian G., Huerta-Sanchez E., Jiang T., Jiang H., Albrechtsen A., Andersen G., Cao H., Korneliusson T., *et al.* (2010) Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat. Genet.*, **42**, 969-972.
7. Bonnefond A., Durand E., Sand O., De G. F., Gallina S., Busiah K., Lobbens S., Simon A., Bellanne-Chantelot C., Letourneau L., *et al.* (2010) Molecular diagnosis of neonatal diabetes mellitus using next-generation sequencing of the whole exome. *PLoS. One.*, **5**, e13630.
8. Li H., Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics.*, **26**, 589-595.
9. McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytzky A., Garimella K., Altshuler D., Gabriel S., Daly M., DePristo M. A. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297-1303.
10. Adamusiak T., Parkinson H., Muilu J., Roos E., van der Velde K. J., Thorisson G. A., Byrne M., Pang C., Gollapudi S., Ferretti V., *et al.* (2012) Observ-OM and Observ-TAB: Universal syntax solutions for the integration, search, and exchange of phenotype and genotype information. *Hum. Mutat.*, **33**, 867-873.
11. Adzhubei I. A., Schmidt S., Peshkin L., Ramensky V. E., Gerasimova A., Bork P., Kondrashov A. S., Sunyaev S. R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248-249.
12. Kumar P., Henikoff S., Ng P. C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073-1081.
13. Mathe E., Olivier M., Kato S., Ishioka C., Hainaut P., Tavtigian S. V. (2006) Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. *Nucleic Acids Res.*, **34**, 1317-1325.
14. Tavtigian S. V., Deffenbaugh A. M., Yin L., Judkins T., Scholl T., Samollow P. B., de S. D., Zharkikh A., Thomas A. (2006) Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J. Med. Genet.*, **43**, 295-305.
15. Gilissen C., Hoischen A., Brunner H. G., Veltman J. A. (2012) Disease gene identification strategies for exome sequencing. *Eur. J. Hum. Genet.*
16. MacArthur D. G., Balasubramanian S., Frankish A., Huang N., Morris J., Walter K., Jostins L., Habegger L., Pickrell J. K., Montgomery S. B., *et al.* (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, **335**, 823-828.
17. Sudmant P. H., Kitzman J. O., Antonacci F., Alkan C., Malig M., Tsalenko A., Sampsas N., Bruhn L., Shendure J., Eichler E. E. (2010) Diversity of human copy number variation and multicopy genes. *Science*, **330**, 641-646.
18. Ng S. B., Nickerson D. A., Bamshad M. J., Shendure J. (2010) Massively parallel sequencing and rare disease. *Hum. Mol. Genet.*, **19**, R119-R124.
19. Nijman I. J., Mokry M., van B. R., Toonen P., de B. E., Cuppen E. (2010) Mutation discovery by targeted genomic enrichment of multiplexed barcoded samples. *Nat. Methods*, **7**, 913-915.
20. Hirschfeld M., Weis J. J., Toshchakov V., Salkowski C. A., Cody M. J., Ward D. C., Qureshi N., Michalek S. M., Vogel S. N. (2001) Signaling by toll-like receptor 2 and 4 agonists results in

- differential gene expression in murine macrophages. *Infect. Immun.*, **69**, 1477-1482.
21. Piccini A., Carta S., Tassi S., Lasiglie D., Fossati G., Rubartelli A. (2008) ATP is released by monocytes stimulated with pathogen-sensing receptor ligands and induces IL-1beta and IL-18 secretion in an autocrine way. *Proc. Natl. Acad. Sci. U. S. A*, **105**, 8067-8072.
  22. Grutter C., Briand C., Capitani G., Mittl P. R., Papin S., Tschopp J., Grutter M. G. (2006) Structure of the PRYSPRY-domain: implications for autoinflammatory diseases. *FEBS Lett.*, **580**, 99-106.
  23. Keeble A. H., Khan Z., Forster A., James L. C. (2008) TRIM21 is an IgG receptor that is structurally, thermodynamically, and kinetically conserved. *Proc. Natl. Acad. Sci. U. S. A*, **105**, 6045-6050.
  24. Woo J. S., Imm J. H., Min C. K., Kim K. J., Cha S. S., Oh B. H. (2006) Structural and functional insights into the B30.2/SPRY domain. *EMBO J.*, **25**, 1353-1363.
  25. Weinert C., Grutter C., Roschitzki-Voser H., Mittl P. R., Grutter M. G. (2009) The crystal structure of human pyrin b30.2 domain: implications for mutations associated with familial Mediterranean fever. *J. Mol. Biol.*, **394**, 226-236.
  26. Chae J. J., Wood G., Richard K., Jaffe H., Colburn N. T., Masters S. L., Gumucio D. L., Shoham N. G., Kastner D. L. (2008) The familial Mediterranean fever protein, pyrin, is cleaved by caspase-1 and activates NF-kappaB through its N-terminal fragment. *Blood*, **112**, 1794-1803.
  27. Diaz A., Hu C., Kastner D. L., Schaner P., Reginato A. M., Richards N., Gumucio D. L. (2004) Lipopolysaccharide-induced expression of multiple alternatively spliced MEFV transcripts in human synovial fibroblasts: a prominent splice isoform lacks the C-terminal domain that is highly mutated in familial Mediterranean fever. *Arthritis Rheum.*, **50**, 3679-3689.
  28. Mansfield E., Chae J. J., Komarow H. D., Brotz T. M., Frucht D. M., Aksentijevich I., Kastner D. L. (2001) The familial Mediterranean fever protein, pyrin, associates with microtubules and colocalizes with actin filaments. *Blood*, **98**, 851-859.
  29. Martinon F., Tschopp J. (2004) Inflammatory caspases: linking an intracellular innate immune system to autoinflammatory diseases. *Cell*, **117**, 561-574.
  30. Chae J. J., Komarow H. D., Cheng J., Wood G., Raben N., Liu P. P., Kastner D. L. (2003) Targeted disruption of pyrin, the FMF protein, causes heightened sensitivity to endotoxin and a defect in macrophage apoptosis. *Mol. Cell*, **11**, 591-604.
  31. Chae J. J., Wood G., Masters S. L., Richard K., Park G., Smith B. J., Kastner D. L. (2006) The B30.2 domain of pyrin, the familial Mediterranean fever protein, interacts directly with caspase-1 to modulate IL-1beta production. *Proc. Natl. Acad. Sci. U. S. A*, **103**, 9982-9987.
  32. Yu J. W., Wu J., Zhang Z., Datta P., Ibrahim I., Taniguchi S., Sagara J., Fernandes-Alnemri T., Alnemri E. S. (2006) Cryopyrin and pyrin activate caspase-1, but not NF-kappaB, via ASC oligomerization. *Cell Death. Differ.*, **13**, 236-249.
  33. Chae J. J., Cho Y. H., Lee G. S., Cheng J., Liu P. P., Feigenbaum L., Katz S. I., Kastner D. L. (2011) Gain-of-function Pyrin mutations induce NLRP3 protein-independent interleukin-1beta activation and severe autoinflammation in mice. *Immunity*, **34**, 755-768.
  34. Hall M. W., Gavrilin M. A., Knatz N. L., Duncan M. D., Fernandez S. A., Wewers M. D. (2007) Monocyte mRNA phenotype and adverse outcomes from pediatric multiple organ dysfunction syndrome. *Pediatr. Res.*, **62**, 597-603.
  35. Papin S., Cuenin S., Agostini L., Martinon F., Werner S., Beer H. D., Grutter C., Grutter M., Tschopp J. (2007) The SPRY domain of Pyrin, mutated in familial Mediterranean fever patients, interacts with inflammasome components and inhibits proIL-1beta processing. *Cell Death. Differ.*, **14**, 1457-1466.
  36. Seshadri S., Duncan M. D., Hart J. M., Gavrilin M. A., Wewers M. D. (2007) Pyrin levels in human monocytes and monocyte-derived macrophages regulate IL-1beta processing and release. *J. Immunol.*, **179**, 1274-1281.
  37. Yuval Y., Hemo-Zisser M., Zemer D., Sohar E., Pras M. (1995) Dominant inheritance in two families with familial Mediterranean fever (FMF). *Am. J. Med. Genet.*, **57**, 455-457.
  38. (1997) Ancient missense mutations in a new member of the RoRet gene family are likely to cause familial Mediterranean fever. The International FMF Consortium. *Cell*, **90**, 797-807.
  39. van Gijn M. E., Soler S., de la Chapelle C., Mulder M., Ritorre C., Kriek M., Philibert L., van der Wielen M., Frenkel J., Grandemange S., et al. (2008) Search for copy number alterations in the MEFV gene using multiplex ligation probe amplification, experience from three diagnostic centres. *Eur. J. Hum. Genet.*, **16**, 1404-1406.

40. Booty M. G., Chae J. J., Masters S. L., Remmers E. F., Barham B., Le J. M., Barron K. S., Holland S. M., Kastner D. L., Aksentjevich I. (2009) Familial Mediterranean fever with a single MEFV mutation: where is the second hit? *Arthritis Rheum.*, **60**, 1851-1861.
41. Singh-Grewal D., Chaitow J., Aksentjevich I., Christodoulou J. (2007) Coexistent MEFV and CIAS1 mutations manifesting as familial Mediterranean fever plus deafness. *Ann. Rheum. Dis.*, **66**, 1541.
42. Stojanov S., Kastner D. L. (2005) Familial autoinflammatory diseases: genetics, pathogenesis and treatment. *Curr. Opin. Rheumatol.*, **17**, 586-599.
43. Touitou I., Perez C., Dumont B., Federici L., Jorgensen C. (2006) Refractory auto-inflammatory syndrome associated with digenic transmission of low-penetrance tumour necrosis factor receptor-associated periodic syndrome and cryopyrin-associated periodic syndrome mutations. *Ann. Rheum. Dis.*, **65**, 1530-1531.
44. Kirecetepe A. K., Kasapcopur O., Arisoy N., Celikyapi E. G., Hatemi G., Ozdogan H., Tahir T. E. (2011) Analysis of MEFV exon methylation and expression patterns in familial Mediterranean fever. *BMC. Med. Genet.*, **12**, 105.
45. Kastner D. L. (1998) Familial Mediterranean fever: the genetics of inflammation. *Hosp. Pract. (Minneap. )*, **33**, 131-40, 143.
46. Booth D. R., Gillmore J. D., Lachmann H. J., Booth S. E., Bybee A., Soyuturk M., Akar S., Pepys M. B., Tunca M., Hawkins P. N. (2000) The genetic basis of autosomal dominant familial Mediterranean fever. *QJM.*, **93**, 217-221.
47. Caglar M. K., Altugan F. S., Ozyurt H., Atasoy H. I. (2008) Screening of family members of children with Familial Mediterranean Fever: true-autosomal and pseudo-autosomal inheritance. *Acta Reumatol. Port.*, **33**, 415-420.
48. Porrello M., Ciaccio C., Gallerano P., Gesu M., Forni G., Mancuso A. (2003) [Familial Mediterranean Fever: a case of dominant transmission with variable penetrance]. *Pediatr. Med. Chir.*, **25**, 289-291.
49. Saito M., Fujisawa A., Nishikomori R., Kambe N., Nakata-Hizume M., Yoshimoto M., Ohmori K., Okafuji I., Yoshioka T., Kusunoki T., *et al.* (2005) Somatic mosaicism of CIAS1 in a patient with chronic infantile neurologic, cutaneous, articular syndrome. *Arthritis Rheum.*, **52**, 3579-3585.
50. Saito M., Nishikomori R., Kambe N., Fujisawa A., Tanizaki H., Takeichi K., Imagawa T., Iehara T., Takada H., Matsubayashi T., *et al.* (2008) Disease-associated CIAS1 mutations induce monocyte death, revealing low-level mosaicism in mutation-negative cryopyrin-associated periodic syndrome patients. *Blood*, **111**, 2132-2141.
51. Tanaka N., Izawa K., Saito M. K., Sakuma M., Oshima K., Ohara O., Nishikomori R., Morimoto T., Kambe N., Goldbach-Mansky R., *et al.* (2011) High incidence of NLRP3 somatic mosaicism in patients with chronic infantile neurologic, cutaneous, articular syndrome: results of an International Multicenter Collaborative Study. *Arthritis Rheum.*, **63**, 3625-3632.
52. Aldea A., Campistol J. M., Arostegui J. I., Rius J., Maso M., Vives J., Yague J. (2004) A severe autosomal-dominant periodic inflammatory disorder with renal AA amyloidosis and colchicine resistance associated to the MEFV H478Y variant in a Spanish kindred: an unusual familial Mediterranean fever phenotype or another MEFV-associated periodic inflammatory disorder? *Am. J. Med. Genet. A*, **124A**, 67-73.

## ***Exome sequencing in a family segregating for celiac disease***

*Agata Szperl <sup>1\*</sup>, Isis Ricaño-Ponce <sup>1\*</sup>, Jiankang Li <sup>2\*</sup>, Patrick Deelen <sup>1</sup>, Alexandros Kanterakis <sup>1</sup>, Vincent Plagnol <sup>3</sup>, Freerk van Dijk <sup>1</sup>, Harm-Jan Westra <sup>1</sup>, Gosia Trynka <sup>1</sup>, Chris J Mulder <sup>4</sup>, Morris Swertz <sup>1</sup>, Wijmenga C <sup>1#</sup> and Hancheng Ch Zheng <sup>2#</sup>*

*\*. # These authors contributed equally*

*<sup>1</sup> Department of Genetics, University Medical Center Groningen and University of Groningen, Groningen, the Netherlands*

*<sup>2</sup> Department of BioMedical Research, Research & Cooperation Division, BGI-Shenzhen, Shenzhen, China*

*<sup>3</sup> UCL Genetics Institute, University College London, London, UK*

*<sup>4</sup> Department of Gastroenterology, VU Medical Center, Amsterdam, the Netherlands*

*This manuscript was published in Clinical Genetics, 2011, 80:*

*138-147.*



## Abstract

Celiac disease is a multifactorial disorder caused by an unknown number of genetic factors interacting with an environmental factor. Hence, most patients are singletons and large families segregating with celiac disease are rare. We report on a three-generation family with six patients in which the inheritance pattern is consistent with an autosomal dominant model. To date, 27 loci explain up to 40% of the heritable disease risk. We hypothesized that part of the missing heritability is due to low frequency- or rare variants. Such causal variants could be more prominent in multi-generation families where private mutations might co-segregate with the disease. They can be identified by linkage analysis combined with whole exome sequencing. We found three linkage regions on 4q32.3-4q33, 8q24.13-8q24.21 and 10q23.1-10q23.32 that segregate with celiac disease in this family. We performed exome sequencing on two affected individuals to investigate the positional candidate regions and the remaining exome for causal nonsense variants. We identified 12 nonsense mutations with a low frequency (MAF < 10%) present in both individuals, but none mapped to the linkage regions. Two variants in the *CSAG1* and *KRT37* genes were present in all six affected individuals. Two nonsense variants in the *MADD* and *GBGT1* genes were also present in 5 of 6 and 4 of 6 individuals, respectively; future studies should determine if any of these nonsense variants is causally related to celiac disease.

## Introduction

Celiac disease (CeD) is a complex, autoimmune disease triggered by dietary gluten, which is widely available in cereals such as barley, rye and wheat. CeD is primarily a T cell-mediated immune disorder, in which CD4+ T cells recognize gluten peptides, resulting in a strong inflammatory response in the small intestine. CeD is a classic example of a multifactorial disease caused by many genetic factors, in addition to the environmental factor. It has been well established that the human leukocyte antigen (HLA) molecules HLA-DQ2 and -DQ8 play a key role in CeD pathogenesis (1, 2). These are also the most important genetic factors associated with the disease and explain some 35% of the heritability. Genome-wide association studies (GWAS) recently identified 26 non-HLA loci that contribute to CeD and explain an additional 5% of the heritability with their modest effect size and odds ratios lower than 1.5. These loci comprise 69 genes that are mainly involved in the immune response (3). Family-based linkage studies may offer a powerful alternative to identifying more CeD genes with a larger effect size. However, there are few large families showing segregation of complex diseases such as CeD. In 2004, Van Belzen et al. (4) reported two linkage regions from a four-generation, Dutch CeD family with 17 affected individuals. Direct sequencing of positional candidate genes from the *9p13-21* region did not reveal causative mutations (5). The lack of high-throughput methods to investigate all the candidate genes from the large linkage regions hampered progress on this family, but work is ongoing. Recently, exome sequencing has been reviewed as a rapid, high-throughput tool for mutation screening (6) and successfully used to identify rare causal mutations, not only in Mendelian diseases (7) but also in complex diseases (8).

We analyzed a second, large CeD family of Caucasian origin with six patients segregating the disease across three generations and with suggested autosomal dominant inheritance. Linkage analysis revealed three potential loci on 4q32.1-q33 (12 Mb), 8q24.13-8q24.21 (5 Mb) and 10q23.1-10q23.32 (10 Mb) chromosomes. We hypothesized that the causal variant responsible for CeD in this family would be a point mutation resulting in a non-functional protein product, and that it would be present in one of the linkage regions. To identify such a mutation, we performed exome sequencing of two affected family members to screen for nonsense mutations in the 148 genes making up the three linkage regions.

## Material and methods

### The study family

The family is Dutch of Caucasian origin and includes six CeD patients, all carrying the HLA-DQ2 genotype (Supplementary Fig. 1, Table 1). The detailed inheritance of the HLA-DQ2 and DQ8 risk genotypes is based on five tagging SNPs from the ImmunoChip that are specific for the HLA alleles present in the Dutch population (9).

In this family, CeD affects approximately 38.5% of the offspring in the second generation. Four out of six of the affected individuals were diagnosed via a small intestine biopsy (Table 1). All the biopsies were re-evaluated and classified by a gastroenterologist (C.J.M.) as Marsh IIIa, IIIb, or IIIc (i.e. partial-, subtotal-, or total villous atrophy, with the presence of crypt hyperplasia and increased number of intraepithelial lymphocytes (30 per 100 enterocytes). Genomic DNA was isolated from peripheral blood to perform genotyping for linkage analysis and exome sequencing.

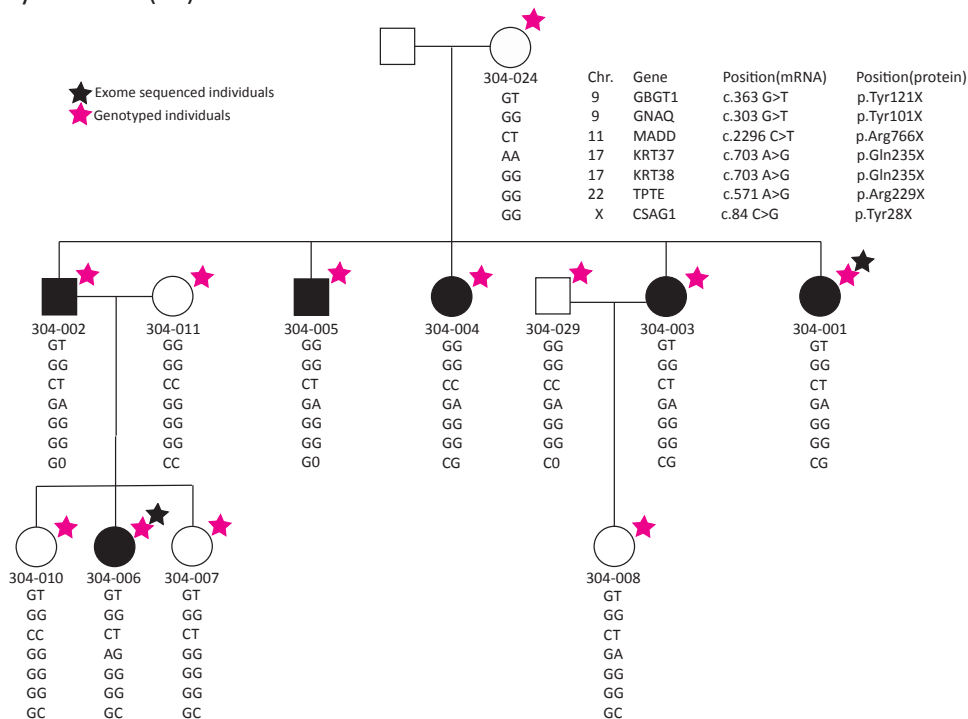
The study was approved by the ethics review board of the University Medical Center Groningen and written informed consent was obtained from all participants.

**Table 1. Characteristics of a three-generation Dutch family affected by celiac disease.**

Individual	Gender	Year of birth	Marsh status	DQ Type	Status for this study
CD0304-001	Female	1954	IIIa	DQ2	Affected
CD0304-002	Male	1951	IIIb	DQ2	Affected
CD0304-003	Female	1952	IIIb	DQ2	Affected
CD0304-004	Female	1948	Undetermined	DQ2	Affected
CD0304-005	Male	1947	IIIb	DQ2	Affected
CD0304-006	Female	1984	Undetermined	DQ2/DQ8	Affected
CD0304-007	Female	1985	Undetermined	DQ2	Unaffected
CD0304-008	Female	1972	Undetermined	DQ2	Unaffected
CD0304-010	Female	1982	Undetermined	DQ2	Unaffected
CD0304-011	Female	1953	Undetermined	DQ8	Unaffected
CD0304-024	Female	1919	Undetermined	DQ2	Unaffected
CD0304-029	Male	1949	Undetermined	DQX	Unaffected

## Genotyping

Six affected (CD0304-001, CD0304-002, CD0304-003, CD0304-004, CD0304-005, CD0304-006,) and six unaffected (CD0304-007, CD0304-010, CD0304-011, CD0304-029, CD0304-008, CD0304-024) family members were genotyped using the ImmunoChip (10) (Fig. 1) as described in Illumina's protocols. The National Center for Biotechnology Information (NCBI) build 36 (hg18) (Illumina manifest file Immuno\_BeadChip\_11419691\_B.bpm) and the second-generation Rutgers combined linkage-physical map were used for mapping (11). The quality control was performed in Plink v1.07 (12). First we checked for any individuals missing more than 5% of the genotypes, but they all had more than 95% of the genotypes called. We removed all SNPs that had a genotype rate below 95%. We also checked for Mendelian errors using default values since we did not allow >5% Mendelian errors for the entire family, and any SNPs with >10% Mendelian error rate were excluded. We found no Mendelian errors. We included 174,624 SNPs in our study. The Dutch case-control data quality control was performed independently, as described in Trynka *et al.* (13).



**Figure 1. Pedigree presenting the “affected only” part of the family studied and the segregation of seven nonsense sequencing variants (SVs). Affected individuals are marked in black.**

## Linkage analysis

For our linkage analysis with the “affected only” approach, we used the data of selected markers from the 196,524 variants available on the ImmunoChip genotyping array for all 12 individuals (Fig. 1). We used Plink v1.07 for marker selection to apply stringent quality control on the ImmunoChip data, based on call rate (> 99%) and Hardy-Weinberg equilibrium ( $p > 0.001$ ). We excluded indels and only considered highly polymorphic SNPs (MAF > 20%). SNPs showing any sign of Mendelian inconsistency were excluded. We then trimmed the dataset to limit biases associated with linkage disequilibrium, pruning SNPs so that each 50-SNP window contained no pair of SNPs with  $r^2 > 0.2$ . Finally we included 8,750 markers for the analysis distributed equally over the genome, providing a sufficiently dense coverage to perform linkage analysis (Supplementary Fig. 2). The linkage information content was uniformly larger than 0.75 (14). Chromosome X was excluded from the linkage analysis as the results are often difficult to interpret. We used the second-generation Rutgers combined linkage-physical map (11) for mapping the 8,750 markers used in our linkage analysis.

Parametric and non-parametric linkage analysis was performed using Merlin v1.1.2 (15). The parametric model assumed a dominant inheritance with a disease probability of 1% for non-carriers and 80% for carriers.

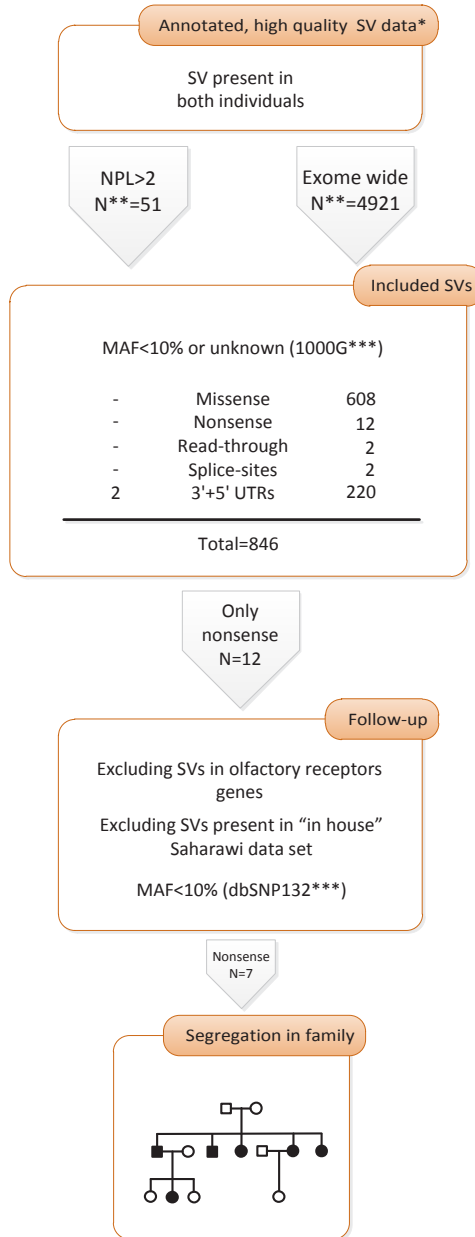
## Exome sequencing: Library generation, reference alignment and variant calling

Library generation, reference alignment and variant calling were performed at BGI, as described in Li et al., 2010 (16). In brief, 5  $\mu\text{g}$  of high quality DNA from two individuals (CD0304-001 and CD0304-006) was fragmented and subsequently hybridized to a NimbleGen 2.1M Human Exome Array. This enrichment captures ~30 Mb of coding DNA, which accounts for approximately 180,000 coding exons. Enriched exome DNA was amplified by PCR followed by random ligation of DNA fragments to the Illumina-compatible adapters and subjected to Solexa library preparation and single-run sequencing of 90 bp per read on average. Before alignment with the reference sequence, low quality reads (containing more than six Ns and/or 40 continuous identical bases and/or polluted by linker or adapter sequences) were removed. SOAPaligner (soap2.20) (17, 18) was used to align clean reads to the human reference genome (NCBI build 36.3) with a maximum of two mismatches allowed. SOAPsnp (19) was used for calling variants in the target region

and with the 500 bp up- and down-stream target regions referred to as “near target region”. We extracted genotypes which differed from the reference as candidate SNPs and kept only sequence variants with a quality score higher than 20, a depth between 4 and 200, an estimated copy number  $\leq 2$ , and a distance between two SNPs of no more than 5, for further analysis. SOAPdenovo was used to identify insertions and deletions (indels) in the exome data by performing a *de novo* assembly of the sequencing reads. Assembled consensus sequences were aligned to the reference genome by a local alignment search tool Z (LASTZ) for aligning two DNA sequences, and inferring appropriate scoring parameters automatically and passed the alignment result to axtBes7t (20) to separate orthologous from paralogous alignment. Finally, we identified the breakpoints in the alignment and annotated the genotypes of the insertions and deletions.

### Annotation and filtration of the sequenced variants

Each of the sequenced variants was annotated for functionality and frequency using the SeattleAnnotation tool (SeattleSeq Annotation, <http://gvs.gs.washington.edu/SeattleSeqAnnotation>), annotate variation ANNOVAR (21) and an in-house pipeline. For MAF annotation, we used the whole genome sequencing of 60 individuals (120 genomes) of European origin from the 1000 Genomes Project (22) (vol1.ftp.pilot\_data.release.2010\_07.low\_coverage). For our analysis we only included sequence variants with a MAF < 10% and only variants present in the exons and splice-sites: i.e. non-synonymous, nonsense, read-through variants and variants in the 3' and 5' UTRs. For the follow-up study, we restricted our analysis to nonsense variants. We excluded nonsense variants present in olfactory genes and having a MAF > 10% in dbSNP132 from the follow-up study (here we considered only the MAF of established Caucasians and based on a greater number of chromosomes than in the 1000 Genomes Project). Nonsense variants present in our data from the exome sequencing of 16 CeD cases from eight Saharawi families (23) were also excluded. Finally, the candidate nonsense SVs were investigated for co-segregation with CeD in the study family. Fig. 2 shows a general flow scheme for our analysis. Indels were annotated using ANNOVAR (21) and sorting intolerant from tolerant (SIFT) (24) and filtrated separately. We considered novel indels (not present in the dbSNP data set) mapping to exonic and splice sites of the genes as interesting (annotation of the gene region was based on the University of California Santa Cruz (UCSC) database (25)).



\* This analysis is only for single nucleotide SVs. Indels were analyzed separately (details are presented in Material&methods section)

\*\* Excluding intragenic and intronic SVs

\*\*\* MAF estimated based on; low coverage 1000 genomes project (120 Europeans genomes) and on the dbSNP132 (only if number of genomes was greater than 120)

**Figure 2. Filtering and follow-up scheme used in our analysis.** We have presented a number of variants that correspond to each step taken.

## Sanger sequencing

We validated variants by direct re-sequencing using a standard Sanger method (26). After filtering, the candidate variants were re-sequenced in the two exome-sequenced individuals (CD0304-001 and CD0304-006) for validation. If the variant was true and present in both individuals, we investigated its segregation in the entire family by re-sequencing it in the other 10 members. Details on the PCR and primers used for validation are summarized in Supplementary table 1.

## Results

### SNP-based linkage analysis

Twelve family members were genotyped on ImmunoChip. After applying quality control, we performed genome-wide, parametric (dominant model) and non-parametric analysis. We found three loci with non-parametric analysis and with a non-parametric LOD score (NPL) of  $\sim 2.40$  on 4q32.3-4q33 ( $p = 0.0004$ ), 8q24.13-8q24.21 ( $p = 0.0004$ ) and 10q23.1-10q23.32 ( $p = 0.0004$ ), which together contained 148 genes (Table 2). Parametric linkage analysis did not identify any more regions with a LOD score  $> 2$ . From non-parametrical and parametrical linkage analysis, we identified six suggestive regions with  $1 < \text{NPL} < 2$ , and four suggestive regions with  $1 < \text{LOD} < 2$  (Supplementary table 2). Some of these regions overlap and others are specific for a single analysis (Table 2, Supplementary table 2).

**Table 2. Three linkage regions with a non-parametric LOD score of  $\sim 2$ .**

Linkage region	Size (Mb)	Number of genes	Model	Non-parametric LOD score	p-value
4q32.3-4q33	5.2	29	Non-parametric	2.40	0.0004
8q24.13-8q24.21	5.0	39	Non-parametric	2.41	0.0004
10q23.1-10q23.32	9.7	80	Non-parametric	2.41	0.0004

### Evaluation of the exome sequencing data

Because of the large number of genes (148) present in the linkage regions, we decided to perform exome sequencing in two affected individuals (CD0304-001 and CD0304-006) (Fig. 1). After enriching for  $\sim 30$  Mb of coding sequence, we obtained on average 2.5 Gb of mapped sequence data per individual. The median exome coverage was 44-fold with 94% of the target region covered with a minimum of 10x (Supplementary table 3). On average, per individual, we identified 18,000



high quality ( $Q > 20$ ) sequence variants in the coding regions (Supplementary table 4). To exclude any possible mix-up of samples and for extra validation of the sequenced data, we compared their genotypes from the ImmunoChip platform with their sequenced variants. We observed a concordance  $> 98\%$  between the two datasets, indicating no sample mix-ups and suggesting a high level of confidence for the sequenced data. We also showed that the concordance between the sequenced individuals was  $\sim 53\%$ , which is in agreement with the genetic distance between the two family members. More detailed statistics of the sequences and SVs can be found in Supplementary tables 3 and 4.

**Table 3. Details of seven nonsense sequence variants shared by two exome-sequenced individuals.**

Gene	Full name of the protein	Chr	Rs number	Genomic	mRNA	Protein
<i>GBGT1</i>	globoside alpha-1,3-N-acetylgalactosaminyl-transferase 1	9	rs35898523	g.135019466	c.363 G>T	p.Tyr121X
<i>GNAQ</i>	Guanine nucleotide-binding protein alpha-q	9		g.79726915	c.303 G>T	p.Tyr101X
<i>MADD</i>	MAP-kinase activating death domain	11	rs35233100	g.47263206	c.2296 C>T	p.Arg766X
<i>KRT37</i>	keratin-37	17		g.36832585	c.703 A>G	p.Gln235X
<i>KRT38</i>	keratin-38	17		g.36849010	c.703 A>G	p.Gln235X
<i>TPTE</i>	transmembrane phosphatase with tensin homology	21	rs1810540	g.9964627	c.571 A>G	p.Arg191X
<i>CSAG1</i>	chondrosarcoma associated gene 1	X	rs1894360	g.151659501	c.84 C>G	p.Tyr28X

### Inclusion filtering of sequence variants

To identify potential disease-causing variants, we adapted a filtering and follow-up scheme that assumed the disease in both family members was due to the same causal variant (Fig. 2). Furthermore, we only included variants with a MAF  $< 10\%$  in the 1000 Genomes Project dataset and having functional effects on the protein product (i.e. missense, splice site, nonsense, read-through variants and SVs from UTRs) (Fig. 2). As the linkage regions are identified with 95% confidence, we could still miss the real disease-causing mutation. Hence, we also applied the same filtering criteria to the entire exome. In total, we identified 846 candidate SVs when we combined the linkage regions (2 SVs) and the exome-wide data (844 SVs) (Fig. 2, Supplementary table 5A). Due to the large number of candidate SVs to be

investigated, we decided to continue the follow-up studies for only the 12 nonsense variations as these are the most damaging (27), although none of them were located in the three linkage regions. We excluded nonsense variants in olfactory genes and those having a MAF > 10% in dbSNP132 and present in our in-house set of Saharawi samples from further analysis (Fig. 2). We also removed the variants in the *CDC27* gene due to the high number of SNPs in the exons; leaving seven nonsense variants to be investigated for co-segregation in the family (Table 3, Supplementary table 6A). The segregation of these variants in the family is shown in Fig. 1. Three variants in the *GNAQ*, *KRT38* and *TPTE* genes were false-positive, meaning that we could not validate these SVs by Sanger sequencing; two variants in the *CSAG1* and *KRT37* genes co-segregate fully with the disease and are present in all the affected individuals; and two other variants are present in the *MADD* and *GBGT1* genes in 5 of 6 and 4 of 6 of the affected individuals, respectively.

To account for the possibility of unequal coverage of sequence in the two family members (CD0304-001 and CD0304-006), we also analyzed them separately and identified an extra 29 and 23 nonsense variants, respectively (Supplementary table 5B). Again, none of these were located in the three linkage regions. Some were present in both individuals and had already been included in the initial analysis, while some were found in only one individual. After applying our final inclusion criteria and discarding one variant in *PRAMEF* gene due to the high number of SNPs in the exons, we investigated 20 SVs that were present in only one individual for validation in the second individual (Supplementary table 6B). None of these were investigated for co-segregation, since 18 could not be confirmed in the second individual, one was false-positive and one, in the *TPTE* gene, we failed to validate with Sanger sequencing due to the presence of small deletions surrounding the candidate SV.

We applied a separate annotation and filtration to the indels, so that they included only novel variants that mapped to the exonic regions and were present in both individuals. In total, we identified 3,258 shared indels, of which 92 were novel (not present in the dbSNP dataset) and mapped to the exons or splice sites of known genes (UCSC was used as a reference). Nine out of 92 were found in the regions of suggestive linkage; two mapped to the exonic regions but did not change the frame, and seven mapped to 3'UTRs (Supplementary table 7), but none were found to be present in the linkage loci of  $NPL > 2$ .

## Discussion

We studied a three-generation Dutch family of Caucasian origin with a dominant-like segregation of CeD to find causative variants that might have a substantial effect on the inherited disease risk. To map the candidate variants, we combined linkage analysis with an “affected only” approach of the entire family with exome sequencing of two affected individuals. As the inheritance model of CeD in the family is uncertain, we also applied a non-parametric linkage approach in addition to the parametric analysis. Both analyses gave comparable results, whereas applying the wrong model for parametric analysis can result in loss of power for detecting linkage (14). We identified three candidate regions using a non-parametric analysis with  $NPL > 2$  but were not able to detect a region with  $LOD > 2$  using parametric analysis. None of the regions overlaps with previously identified CeD loci (3). We did not observe linkage to the HLA region on 6p21 in this family, despite the fact that the HLA-DQ2 and DQ8 loci are the strongest risk factors contributing to CeD. Because HLA-DQ2 and DQ8 alleles are also very common in the general population (~30%), the risk alleles were also inherited from an unaffected parent who married into the family (Supplementary Fig. 1), thereby disrupting the proper segregation of alleles in the family. Hence, linkage to HLA was not identified by our linkage analysis.

CeD is genetically heterogeneous, like other complex diseases, and even within a single family where the inheritance of the disease is compatible with a dominant model, it is likely that multiple loci co-segregate with the disease. Our detection of three linkage regions in this family is in line with a previous linkage analysis in a CeD family in which two regions were found to segregate significantly with CeD (4). The regions found in both families do not overlap, which may also indicate a high heterogeneity for CeD.

In order to identify causal variants, we hypothesized that these variants would be present in the candidate linkage regions and shared by both the affected family members we sequenced. As the three candidate linkage regions together contain 148 genes, exome sequencing is an efficient method to screen all the positional candidate genes for disease-causing variants. At the same time, this technology also allowed us to scrutinize the remainder of the genome in case our linkage analysis proved incorrect. We also hypothesized that within our multi-generation family a limited number of causal variants with substantial risk would be

present. We therefore focused our analysis on nonsense variants as these are the most damaging (27).

After applying our filter criteria to the linkage regions and to the entire exome, we were left with seven nonsense variants, none of which mapped to the linkage regions. After validation by Sanger sequencing, three of them were found to be false-positive, in the *GNAQ*, *KRT38* and *TPTE* genes. The remaining four SVs were investigated for co-segregation with the disease. We were looking for a dominant-like inheritance, but we kept in mind that a low-frequency, causative variant for a complex disease does not have to demonstrate Mendelian segregation but can still contribute to the heritability of the disease (28). Two SVs in the *CSAG1* gene (p.Tyr28X) and *KRT37* (p. Gln235X) were present in all six affected individuals. Regions with these variants were not identified in the linkage analysis because the *CSAG1* gene lies on the X chromosome, which was not submitted for linkage analysis, as it is so difficult to analyze. A variant in the *KRT37* gene was also present in the unaffected spouse that contributed to the “affected only” approach and was thus not picked up in the linkage analysis. The presence of this variant in unaffected spouses could indicate a higher MAF than found in the 1000 Genomes data set, thus case-control genotyping is required for future follow-up studies. Neither gene has an immune-related function: the *CSAG1* gene is reported as a cancer/testis antigen highly expressed in cancer tissues (29), whereas the *KRT37* gene belongs to the type I keratin gene family and is involved in the hair follicle and expressed in epithelial cells (30).

A nonsense variant in the *MADD* gene (p.Arg766X) was present in 5 of 6 affected family members, and a nonsense variant in the *GBGT1* gene (p.Tyr121X) was found in 4 of 6. Thus, neither of these variants segregates fully with the disease, but, interestingly, both of the variants were present on the ImmunoChip which was recently used for CeD case-control studies in 2,312 individuals of Dutch origin. There was no significant association found as the p values associated with these variants were  $p = 0.80$  (MAF=0.059) for *MADD2* and  $p = 0.78$  (MAF=0.079) for *GBGT1* in the Dutch cohort. The *MADD* gene was also found to be associated to type 2 diabetes (31) and it interacts with TNFR1 to activate MAPK and propagate apoptotic signals (32). The *GBGT1* gene is a member of the ABO family that may be involved in tropism and the binding of pathogenic organisms (33).

As the coverage of some regions in our sequencing data could be unequal, we also investigated the nonsense variants in both samples separately. 20 SVs that

were present in only one of the sequenced individuals were investigated further: 18 of 20 genotypes identified in the exome data agreed with the genotypes found in the independent validation step. The variant in the *ZNF81* gene was from a different genotype than in the exome for individual CD0304-006 (Supplementary table 5). The coverage in this region was very low (~2 on average) and this could explain the false-positive calling. We were not able to validate a SV in the *TPTE* gene with Sanger sequencing due to the presence of several deletions surrounding the candidate variant. As none of the remaining true SVs were present in both patients, we did not study them for segregation in the family. In summary, we identified seven nonsense SVs that were shared by the two exome-sequenced patients, but none were located in the linkage regions.

There were a number of weaknesses in our study. First, we assumed that both the affected and sequenced individuals shared the same causal variant, but given the observation of multiple linkage regions in families segregating for complex diseases, this assumption might not be valid. Second, we assumed that the causal variants segregating in a multi-generation family would be nonsense variants, but this might be too stringent and other type of variants possibly influencing protein expression could be followed up in future studies. Risch (34) proposed that while looking for variants causing complex diseases, we should focus on non-synonymous, coding, and 3' and 5' UTRs variants. If we had concentrated on these categories, we would have had 846 variants (844 exome-wide and two from linkage regions) to study further. The two variants identified in the linkage regions mapped to the untranslated regions of two genes: *TLL1* and *SLC16A12*. The variant in *SLC16A12* might be a private variant as it has not been reported in any of the public databases. We investigated the co-segregation of these variants in the family. Minor alleles of the variant in the *TLL1* gene were present in 5 of the 6 affected family members and in 6 of the 6 affected family members in the *SLC16A12* gene (Supplementary Fig. 3). Using the Patrocles algorithm (35), we could verify the consequences of the change on miRNA (miRNA)-binding sites. We observed that variants introduce new miRNA binding sites, which may have functional consequences. However, more extensive case-control and functional studies are needed to prove their potential involvement in the pathogenesis of celiac disease.

From the 844 SVs identified, 27 mapped to the 10 regions of suggestive linkage (NPL and/or LOD > 1 and < 2). Interestingly, the region on chromosome

8q24.13-8q24.21 was also identified as suggestive in parametrical analysis. As recently proposed in a review by Cirulli and Goldstein (6), the investigation of suggestive linkage regions combined with the prioritization of candidate genes is also a way to narrow down the causative variant. We could therefore use this knowledge for future studies in this family.

It seems that an attractive approach might also be to follow up the missense variants with a MAF < 5% (27), however, we would still be left with a large dataset of 502 missense SVs.

We have found 92 exonic, novel indels present in both individuals exome-wide. None of these mapped to the linkage regions with NPL > 2 and nine were found to be present in suggestive linkage regions: seven in the 3' UTR regions and two in protein-coding regions, but not disturbing the protein frame. Because none of these indels were very strong candidates and changes in the UTRs are difficult to interpret, we did not follow up any of these indels.

From recent case-control studies we know that there is an excess of rare missense and nonsense variants in GWAS regions for complex diseases (36, 37). However, none of the seven nonsense variants that we identified mapped to loci previously associated with CeD. Of the 846 SVs, only one mapped to a CeD locus: a missense variant (p.Lys1385Asn) in the *LRRC37A2* gene. Apparently the function of this gene is not well established.

Finally, is possible that the approach we have taken is not appropriate for complex diseases. It might well be that sequencing two individuals is not enough or that the category of variants to focus on should be much broader. A review by Bodmer and Bonilla in 2008 (38) stated that familial-based studies in complex diseases will not have a significant role in finding either rare or common variants because of their low penetrance. If that turns out to be the case, we should focus on those genes or loci that have already been identified by GWAS and perform gene-burden association studies for rare variants in very large case-control studies (36).

In conclusion, although we found three linkage regions that segregate with celiac disease in this family, the approach we chose might not have been suitable for finding the causative variant in this family. We could have missed the causal variant(s) simply because our enrichment covered only 30 Mb of the known and expressed part of the genome. Current exome-capturing kits cover around 50 Mb. It is also possible that true causal variants might be found in the non-coding regions,

as suggested by the large number of observed eQTLs for CeD (3), in which case whole-genome sequencing rather than a whole-exome approach would be more appropriate. Finally, the CeD in this family might be much more complex than we imagined and the occurrence of many patients with CeD, in general, could be more due to chance than to true co-segregation of “serious” causal variants.

## Acknowledgements

The study was supported by a grant from the Netherlands Organization for Scientific Research (NWO, VICI grant 918.66.620 to CW). We thank Jackie Senior for editing this manuscript, Jihane Romanos for helping with analysis and Mathieu Platteel for helping with sample preparation.

## References

1. Sollid L. M., Markussen G., Ek J., Gjerde H., Vartdal F., Thorsby E. (1989) Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/beta heterodimer. *J. Exp. Med.*, **169**, 345-350.
2. Sollid L. M., Thorsby E. (1993) HLA susceptibility genes in celiac disease: genetic mapping and role in pathogenesis. *Gastroenterology*, **105**, 910-922.
3. Dubois P. C., Trynka G., Franke L., Hunt K. A., Romanos J., Curtotti A., Zhernakova A., Heap G. A., Adany R., Aromaa A., *et al.* (2010) Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.*, **42**, 295-302.
4. van Belzen M. J., Vrolijk M. M., Meijer J. W., Crusius J. B., Pearson P. L., Sandkuijl L. A., Houwen R. H., Wijmenga C. (2004) A genomewide screen in a four-generation Dutch family with celiac disease: evidence for linkage to chromosomes 6 and 9. *Am. J. Gastroenterol.*, **99**, 466-471.
5. Wapenaar M. C., Monsuur A. J., Poell J., van 't S. R., Meijer J. W., Meijer G. A., Mulder C. J., Mearin M. L., Wijmenga C. (2007) The SPINK gene family and celiac disease susceptibility. *Immunogenetics*, **59**, 349-357.
6. Cirulli E. T., Goldstein D. B. (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.*, **11**, 415-425.
7. Ng S. B., Buckingham K. J., Lee C., Bigham A. W., Tabor H. K., Dent K. M., Huff C. D., Shannon P. T., Jabs E. W., Nickerson D. A., *et al.* (2010) Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.*, **42**, 30-35.
8. Musunuru K., Pirruccello J. P., Do R., Peloso G. M., Guiducci C., Sougnez C., Garimella K. V., Fisher S., Abreu J., Barry A. J., *et al.* (2010) Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. *N. Engl. J. Med.*, **363**, 2220-2227.
9. Monsuur A. J., de Bakker P. I., Zhernakova A., Pinto D., Verduijn W., Romanos J., Auricchio R., Lopez A., van Heel D. A., Crusius J. B., Wijmenga C. (2008) Effective detection of human leukocyte antigen risk alleles in celiac disease using tag single nucleotide polymorphisms. *PLoS. One.*, **3**, e2270.
10. Cortes A., Brown M. A. (2011) Promise and pitfalls of the Immunochip. *Arthritis Res. Ther.*, **13**, 101.
11. Matise T. C., Chen F., Chen W., De la Vega F. M., Hansen M., He C., Hyland F. C., Kennedy G. C., Kong X., Murray S. S., *et al.* (2007) A second-generation combined linkage physical map of the human genome. *Genome Res.*, **17**, 1783-1786.
12. Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M. A., Bender D., Maller J., Sklar P., de Bakker P. I., Daly M. J., Sham P. C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559-575.

13. Trynka G., Hunt K. A., Bockett N. A., Romanos J., Mistry V., Szperl A., Bakker S. F., Bardella M. T., Bhaw-Rosun L., Castillejo G., *et al.* (2011) Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.*, **43**, 1193-1201.
14. Kruglyak L., Daly M. J., Reeve-Daly M. P., Lander E. S. (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.*, **58**, 1347-1363.
15. Abecasis G. R., Cherny S. S., Cookson W. O., Cardon L. R. (2002) Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.*, **30**, 97-101.
16. Li Y., Vinckenbosch N., Tian G., Huerta-Sanchez E., Jiang T., Jiang H., Albrechtsen A., Andersen G., Cao H., Korneliussen T., *et al.* (2010) Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat. Genet.*, **42**, 969-972.
17. Li R., Li Y., Kristiansen K., Wang J. (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics.*, **24**, 713-714.
18. Li R., Yu C., Li Y., Lam T. W., Yiu S. M., Kristiansen K., Wang J. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics.*, **25**, 1966-1967.
19. Li R., Li Y., Fang X., Yang H., Wang J., Kristiansen K., Wang J. (2009) SNP detection for massively parallel whole-genome resequencing. *Genome Res.*, **19**, 1124-1132.
20. Schwartz S., Kent W. J., Smit A., Zhang Z., Baertsch R., Hardison R. C., Haussler D., Miller W. (2003) Human-mouse alignments with BLASTZ. *Genome Res.*, **13**, 103-107.
21. Wang K., Li M., Hakonarson H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.
22. 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061-1073.
23. Catassi C., Ratsch I. M., Gandolfi L., Pratesi R., Fabiani E., El A. R., Frijia M., Bearzi I., Vizzoni L. (1999) Why is coeliac disease endemic in the people of the Sahara? *Lancet*, **354**, 647-648.
24. Kumar P., Henikoff S., Ng P. C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073-1081.
25. Fujita P. A., Rhead B., Zweig A. S., Hinrichs A. S., Karolchik D., Cline M. S., Goldman M., Barber G. P., Clawson H., Coelho A., *et al.* (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.*, **39**, D876-D882.
26. Szperl A. M., Golachowska M. R., Bruinenberg M., Prekeris R., Thunnissen A. M., Karrenbeld A., Dijkstra G., Hoekstra D., Mercer D., Ksiazek J., *et al.* (2011) Functional characterization of mutations in the myosin Vb gene associated with microvillus inclusion disease. *J. Pediatr. Gastroenterol. Nutr.*, **52**, 307-313.
27. Kryukov G. V., Pennacchio L. A., Sunyaev S. R. (2007) Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.*, **80**, 727-739.
28. Manolio T. A., Collins F. S., Cox N. J., Goldstein D. B., Hindorf L. A., Hunter D. J., McCarthy M. I., Ramos E. M., Cardon L. R., Chakravarti A., *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747-753.
29. Lin C., Mak S., Meitner P. A., Wolf J. M., Bluman E. M., Block J. A., Terek R. M. (2002) Cancer/testis antigen CSAGE is concurrently expressed with MAGE in chondrosarcoma. *Gene*, **285**, 269-278.
30. Rogers M. A., Winter H., Langbein L., Bleiler R., Schweizer J. (2004) The human type I keratin gene family: characterization of new hair follicle specific members and evaluation of the chromosome 17q21.2 gene domain. *Differentiation*, **72**, 527-540.
31. Dupuis J., Langenberg C., Prokopenko I., Saxena R., Soranzo N., Jackson A. U., Wheeler E., Glazer N. L., Bouatia-Naji N., Gloyn A. L., *et al.* (2010) New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.*, **42**, 105-116.
32. Schievella A. R., Chen J. H., Graham J. R., Lin L. L. (1997) MADD, a novel death domain protein that interacts with the type 1 tumor necrosis factor receptor and activates mitogen-activated protein kinase. *J. Biol. Chem.*, **272**, 12069-12075.
33. Xu H., Storch T., Yu M., Elliott S. P., Haslam D. B. (1999) Characterization of the human Forssman synthetase gene. An evolving association between glycolipid synthesis and host-microbial interactions. *J. Biol. Chem.*, **274**, 29390-29398.
34. Risch N. J. (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847-856.



35. Hiard S., Charlier C., Coppieters W., Georges M., Baurain D. (2010) Patrocles: a database of polymorphic miRNA-mediated gene regulation in vertebrates. *Nucleic Acids Res.*, **38**, D640-D651.
36. Johansen C. T., Wang J., Lanktree M. B., Cao H., McIntyre A. D., Ban M. R., Martins R. A., Kennedy B. A., Hassell R. G., Visser M. E., *et al.* (2010) Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat. Genet.*, **42**, 684-687.
37. Rivas M. A., Beaudoin M., Gardet A., Stevens C., Sharma Y., Zhang C. K., Boucher G., Ripke S., Ellinghaus D., Burt N., *et al.* (2011) Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.*, **43**, 1066-1073.
38. Bodmer W., Bonilla C. (2008) Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.*, **40**, 695-701.

***Functional polymorphism in IL12B promoter site is  
associated with ulcerative colitis***

*Agata Szperl<sup>1</sup>, Päivi Saavalainen<sup>2,3</sup>, Rinse-K. Weersma<sup>4</sup>, Maarit Lappalainen<sup>2,5</sup>, Paulina Paavola-Sakki<sup>5,6</sup>, Leena Halme<sup>7</sup>, Martti Färkkilä<sup>6</sup>, Ulla Turunen<sup>6</sup>, Kimmo Kontula<sup>2,5</sup>, Cyriel Y. Ponsioen<sup>8</sup>, Cisca Wijmenga<sup>1</sup>, Cleo C. van Diemen<sup>1</sup>*

*<sup>1</sup> Department of Genetics, University Medical Centre Groningen, University of Groningen, Groningen, the Netherlands*

*<sup>2</sup>Research Program for Molecular Medicine, University of Helsinki, Helsinki, Finland*

*<sup>3</sup>Department of Medical Genetics, Haartman Institute, University of Helsinki, Finland*

*<sup>4</sup> Department of Gastroenterology and Hepatology , University Medical Centre Groningen, University of Groningen, Groningen, the Netherlands*

*<sup>5</sup>Department of Medicine, Helsinki University Hospital, Helsinki, Finland*

*<sup>6</sup>Department of Gastroenterology, Helsinki University Hospital, Helsinki, Finland*

*<sup>7</sup>Department of Surgery, Helsinki University Hospital, Helsinki, Finland*

*<sup>8</sup>Department of Gastroenterology and Hepatology, Academic Medical Center, Amsterdam, the Netherlands*

*This manuscript was published in Journal of Inflammatory Bowel Disease, 2011, 17(6): E38-40*



**To the editor,**

Ulcerative colitis (UC) is one of the two subtypes of inflammatory bowel disease (IBD). UC is a complex genetic disease that involves an abnormal immune response to the intestinal microflora causing ulceration of the epithelial barrier of the colon. UC symptoms include bloody diarrhoea, anaemia and abdominal pain. Twin studies and clustering of IBD in families indicate the presence of genetic factors that account for the development of the disease. Genome-wide association studies (GWAS) for UC have identified 30 loci with mostly immune-related genes such as *IL23R*, *JAK2*, *STAT3* and *IL12B* (1). However, these common variants explain only part of the heritability due to their low effect size (odds ratios < 2). Part of the missing heritability may thus lie in rare causative variants with a larger effect size.

One of the UC associated loci is the *IL12B* gene, encoding the IL12p40 homodimer protein produced by monocytes and dendritic cells. IL12p40 is the common subunit of two interleukins; IL12 and IL23. IL12 is responsible for T helper (Th) 1 differentiation and induces the production of interferon- $\gamma$  (IFN- $\gamma$ ), whereas IL23 is responsible for stabilizing the Th17 phenotype by promoting IL-17 production. A common GWAS associated SNP (rs10045431) in *IL12B* was not associated to UC in a Dutch population (2). Hence, we hypothesized that other variants in the *IL12B* gene that were not detected in the replication study due to their low frequency (rare variants) or simply because they were not tagged by the replication SNP (due to low linkage disequilibrium, LD) might be causative for UC. As expression of cytokines is likely to be regulated by variants in their regulatory and coding sequence, we genotyped a functional *IL12B* promoter variant previously associated with two other auto-immune diseases, systemic lupus erythematosus (SLE) and asthma (3, 4). Further, to search for rare variants, we sequenced the coding part (including splice-sites) of the *IL12B* gene in 350 UC cases of Dutch origin.

Table 1 shows the genotype distribution, allele frequency and statistics of the *IL12B* promoter polymorphism, an insertion/deletion of 4 bp (rs41292470, previously reported as rs17860508), in both a Dutch and a Finnish cohort and in a combined analysis. We found moderate association of the GC allele with UC in Dutch (OR=0.86; 95% CI: 0.76 – 1.31, P=0.025) whereas in the Finish population we observed borderline significant association in the same direction (Table 1). The signal became more significant in a combined analysis (OR=0.85; 95% CI: 0.76 – 0.95, P=0.003). In the recessive model of inheritance GC/GC genotype was associated

with UC in Dutch (OR= 0.77; 95% CI: 0.63-0.93, P=0.008) and combined analysis (OR= 0.79; 95% CI: 0.67-0.94, P=0.008). The promoter variant did not appear to be in LD with the previously reported GWAS SNP (pair-wise LD statistics (PLINK v1.07,  $r^2$  0.037 and  $D'$  0.285), this may explain why this association signal was not identified in our earlier replication study (2).

Different genotypes of the *IL12B* promoter polymorphism have been linked with a broad spectrum of phenotypes. The CTCTAA/CTCTAA insertion genotype has been associated with increased mortality in children with cerebral malaria (5). Recently, the GC/GC genotype was associated to susceptibility for SLE in a Bulgarian cohort (3), whereas the heterozygous genotype has been correlated with severity of asthma in children (5). These studies and the present one indicate that genetic variation in *IL12B* plays an important role in the immune response in many diseases and the genotype of the promoter variant may influence different phenotypes.

The *IL12B* promoter genotype has also been associated with differences in IL12 production by stimulated dendritic cells and higher transcriptional activity of *IL12B* (6, 7). In contrast to previous studies, we did not see any correlation between the *IL12B* genotype and IL12 peripheral blood mRNA expression in 350 Dutch controls (supplementary figure 1). In this new study, we used unstimulated cells so it is possible that the IL12b mRNA levels were not high enough to identify differential expression. We cannot exclude the possibility of measuring different isoforms from the one measured in previous studies. This might happen due to different platforms used for the mRNA quantification.

In search of causative rare variants for UC, we sequenced the coding parts, including splice sites, of *IL12B* in 350 Dutch UC cases. We identified four known SNPs (two intronic and two non-synonymous, all reported in the dbSNP database) and two heterozygous, unknown variants (one nonsense and one splice-site) (supplementary figure 2). We have genotyped all the variants in our Dutch UC cohort and controls and found very low frequencies of the variants exclusively in UC cases (supplementary figure 2). To investigate if there is a significant enrichment of rare variants in the UC Dutch cohort, we performed a pooled association test (Fisher's exact test) for rare, potentially functional polymorphisms: the two unknown splice-site and nonsense variants, and the two known non-synonymous SNPs rs55661460 and rs3213119. We found no significant association at the allele or genotype level. It is possible that we have not genotyped enough individuals and thus do not have enough power to find the association.

**Table 1. Statistical analysis for genotype and allele frequencies of the promoter polymorphism in ulcerative colitis (UC) patients and controls from the Netherlands and Finland.** For the association test of the *IL12B* promoter variant with UC we used  $\chi^2$  statistics. No genotypes for either cases or controls deviated from Hardy-Weinberg equilibrium. <sup>a</sup> *IL12b* promoter variant was genotyped on 5 ng DNA using TaqMan chemistry and custom assays by Applied Biosystems (Applied Biosystems, FosterCity, CA, USA, [www.appliedbiosystems.com](http://www.appliedbiosystems.com)) using a standard protocol provided by them. The PCR assays and allelic discrimination were run using an ABI PRISM 7900HT Sequence Detection System instrument (Applied Biosystems). Data was analyzed using the SDS program 2.3 (Applied Biosystems). <sup>b</sup> *IL12b* promoter variant was genotyped by fragment analysis. Details are provided at <http://seqlab.gju.u.fj/FragmentAnalysis.htm>

	Dutch <sup>a</sup>				Finns <sup>b</sup>				Combined			
	Frequencies		P value	Odds Ratio (CI)	Frequencies		P value	Odds Ratio (CI)	Frequencies		P value	Odds Ratio (CI)
	Cases N=940	Controls N=1085			Cases N=345	Controls N=312			Cases N=1285	Controls N=1397		
<b>Genotypes</b>												
CTCTAA/ CTCTAA	0.22	0.20	<b>0.03</b>	-	0.27	0.18	0.052	-	0.23	0.20	<b>0.013</b>	-
CTCTAA/ GC	0.52	0.48			0.46	0.52			0.51	0.49		
GC/GC	0.26	0.32			0.27	0.30			0.26	0.31		
<b>Alleles</b>												
CTCTAA	0.48	0.44		0.86 (0.76-1.31)	0.50	0.44		0.8 (0.64-1.01)	0.48	0.44		0.85 (0.76-0.95)
GC	0.52	0.56	<b>0.025</b>		0.50	0.56	0.06		0.52	0.56	<b>0.003</b>	
<b>Dominant model for GC allele and recessive for CTCTAA allele</b>												
CTCTAA/ CTCTAA	0.22	0.20	0.35		0.27	0.18	0.15		0.23	0.20	<b>0.04</b>	
CTCTAA/ GC + GC/ GC	0.78	0.80		0.90 (0.72-1.12)	0.73	0.81		0.62 (0.42-0.92)	0.77	0.80		0.82 (0.68-0.99)
<b>Recessive model for GC allele and dominant for CTCTAA allele</b>												
CTCTAA/ CTCTAA + CTCTAA/ GC	0.74	0.68	<b>0.008</b>	0.77 (0.63-0.93)	0.73	0.70	0.48	0.88	0.74	0.69	<b>0.008</b>	0.79 (0.67-0.94)
GC/GC	0.26	0.32			0.27	0.30		(0.62-1.25)	0.26	0.31		

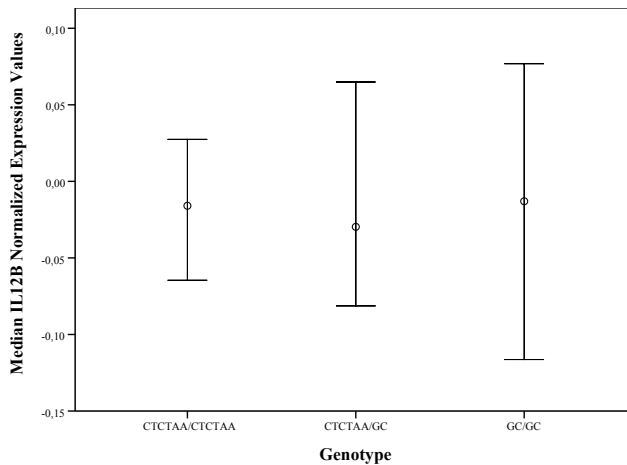
The nonsense mutation reported here (with a frequency of less than 0.05 in Dutch UC patients) maps to the functional domain of the *IL12B* protein and might influence the functionality of the protein (supplementary figure 2). The splice-site variant maps to a non-conserved site in the splice acceptor site.

Overall, we found both a very low number of variants, with low frequencies, in the Dutch UC cohort (supplementary figure 2), in accordance with Huang et al. (8) who sequenced the *IL12B* gene in 72 individuals and did not find any changes that would lead to amino acid substitutions. This suggests that *IL12B* is a key cytokine in which genetic variation is not tolerated. This is further supported by the high conservation of *IL12B* across different species and that *IL12B* is not under evolutionary selection (average integrated haplotype score (iHS) of 0.54 in CEU (Utah residents with European ancestry) GWAS SNP rs3212227 iHS -0.20, analyzed in 200 Dutch controls) (Haplotter: <http://haplotter.uchicago.edu/>).

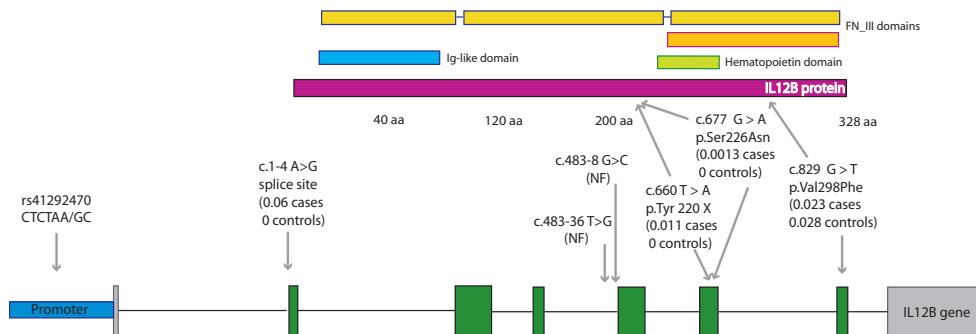
In conclusion, we have shown association of the functional variant in the *IL12B* promoter site with UC. This variant was not tagged by the common variant used for replication of the GWAS was thus not picked up in an earlier Dutch cohort study. We also identified very rare, nonsense mutations that may influence *IL12B* function. This study shows that although some loci are not replicated, they may still be causative and it is therefore important to investigate such loci further.

### Acknowledgements

The study was supported by grand the Netherlands Organization for Scientific Research (NWO-VICI grant 918.66.620 to CW) and by EU Marie Curie Excellence Grant (MEXT-CT-2005-025270. RW is supported by a clinical fellow grant (90700281) from the Netherlands Organization for Scientific Research (NWO). We would like to thank Elisabet Einarsdottir for technical and professional support in Helsinki, Alexandra Zhernakova for helping with the analysis, Anna-Elina Lehesjoki and Albert de la Chapelle for providing the study with Finnish population control DNA samples and Jackie Senior for critically reading this manuscript.



**Supplementary figure 1** Peripheral blood mRNA was collected from 450 control individuals using PAXgene tubes (Qiagen). Starting with 200 ng of RNA, the Ambion Illumina TotalPrep Amplification Kit was used for anti-sense RNA synthesis, amplification, and purification, according to the manufacturer's protocol (Applied Biosystems/Ambion, Austin, TX, USA). 750 ng of complementary RNA was hybridized to Illumina HumanHT12 BeadChips (Illumina, San Diego, CA, USA) and scanned on the Illumina BeadArray Reader. Data were quantile-quantile normalized per tissue using Genespring GX software (Agilent technologies). Differences in IL12B gene expression according to promoter genotype were calculated using one-way ANOVA.



**Supplementary figure 2.** Variants identified through Sanger sequencing (\*) and further genotyped (\*\*) in the Dutch (750 cases, 750 controls\*\*\*) and Finnish (318 cases, 299 controls\*\*\*) cohorts. The protein and genomic location is shown for each variant. The frequencies for the variants in the Dutch cases and controls are shown in brackets. These variants were not found in the Finnish cohort.

NF: variant was not followed by genotyping.

(\*) Sanger sequencing: The PCR reaction was performed with 50 ng genomic DNA. The primers used and conditions of the PCR reaction are available upon request. Sequences were read on a 3730 DNA analyzer and 3130 Genetic analyzer (Applied Biosystems, Foster City, CA, USA; www.appliedbiosystems.com). Sequences were analyzed using ContigExpress software (Invitrogen, Carlsbad, CA) and GeneMapper 4.0 (Applied Biosystems).

(\*\*) Variants were genotyped on 5 ng DNA using TaqMan chemistry assays from Applied Biosystems using a standard protocol provided by them. The PCR assays and allelic discrimination were run using an ABI PRISM 7900HT Sequence Detection System instrument (Applied Biosystems). Data was analyzed using the SDS program 2.3 (Applied Biosystems). DNA was isolated from peripheral blood samples using standard laboratory procedures. Concentrations and purity were determined with NanoDrop ND-1000 (Isogen Life Science, De Meern, the Netherlands).

(\*\*\*) Frequencies were counted for the numbers of individuals after removing drops-out.



## References

1. McGovern D. P., Gardet A., Torkvist L., Goyette P., Essers J., Taylor K. D., Neale B. M., Ong R. T., Lagace C., Li C., *et al.* (2010) Genome-wide association identifies multiple ulcerative colitis susceptibility loci. *Nat. Genet.*, **42**, 332-337.
2. Festen E. A., Stokkers P. C., van Diemen C. C., van Bodegraven A. A., Boezen H. M., Crusius B. J., Hommes D. W., van der Woude C. J., Balschun T., Verspaget H. W., *et al.* (2010) Genetic analysis in a Dutch study sample identifies more ulcerative colitis susceptibility loci and shows their additive role in disease risk. *Am. J. Gastroenterol.*, **105**, 395-402.
3. Miteva L. D., Manolova I. M., Ivanova M. G., Rashkov R. K., Stoilov R. M., Gulubova M. V., Stanilova S. A. (2010) Functional genetic polymorphisms in interleukin-12B gene in association with systemic lupus erythematosus. *Rheumatol. Int.*
4. Morahan G., Huang D., Wu M., Holt B. J., White G. P., Kendall G. E., Sly P. D., Holt P. G. (2002) Association of IL12B promoter polymorphism with severity of atopic and non-atopic asthma in children. *Lancet*, **360**, 455-459.
5. Morahan G., Boutlis C. S., Huang D., Pain A., Saunders J. R., Hobbs M. R., Granger D. L., Weinberg J. B., Peshu N., Mwaikambo E. D., *et al.* (2002) A promoter polymorphism in the gene encoding interleukin-12 p40 (IL12B) is associated with mortality from cerebral malaria and with reduced nitric oxide production. *Genes Immun.*, **3**, 414-418.
6. Muller-Berghaus J., Olson W. C., Moulton R. A., Knapp W. T., Schadendorf D., Storkus W. J. (2005) IL-12 production by human monocyte-derived dendritic cells: looking at the single cell. *J. Immunother.*, **28**, 306-313.
7. Hirota T., Suzuki Y., Hasegawa K., Obara K., Matsuda A., Akaoshi M., Nakashima K., Cheng L., Takahashi N., Shimizu M., *et al.* (2005) Functional haplotypes of IL-12B are associated with childhood atopic asthma. *J. Allergy Clin. Immunol.*, **116**, 789-795.
8. Huang D., Cancilla M. R., Morahan G. (2000) Complete primary structure, chromosomal localisation, and definition of polymorphisms of the gene encoding the human interleukin-12 p40 subunit. *Genes Immun.*, **1**, 515-520.

*Cross-ethnic replication and fine-mapping  
of coeliac disease loci in north Indian population*

*Sabyasachi Senapati<sup>1,\*</sup>, Gosia Trynka<sup>2,\*</sup>, Agata Szperł, Jihane Romanos<sup>2</sup>, Alexandra Zhernakova<sup>2,3,4</sup>, Ajit Sood<sup>5</sup>, Vandana Midha<sup>5</sup>, Andre de Vries<sup>2</sup>, Marcel Kempenaar<sup>2</sup>, Lude Franke<sup>2</sup>, Santos Alonso<sup>6</sup>, Gerard te Meerman<sup>2</sup>, Thelma BK<sup>1,#</sup>, Cisca Wijmenga<sup>2,#</sup>*

*\*,# These authors contributed equally*

*<sup>1</sup>Department of Genetics, University of Delhi, South Campus, New Delhi, India*

*<sup>2</sup>Genetics Department, University of Groningen, University Medical Centre Groningen, Groningen, the Netherlands*

*<sup>3</sup>Department of Rheumatology, Leiden University Medical Centre, Leiden, the Netherlands*

*<sup>4</sup>Complex Genetics Section, Department of Biomedical Genetics, University Medical Centre Utrecht, Utrecht, the Netherlands*

*<sup>5</sup>Dayanand Medical College and Hospital, Ludhiana, Punjab, India*

*<sup>6</sup>Department of Genetics, Physical Anthropology and Animal Physiology, University of the Basque Country, Leioa, Bizkaia, Spain*

## Abstract

Genome-wide association studies (GWAS) in coeliac disease have shown strong and replicable association to 26 non-HLA loci in individuals of European descent. Since coeliac disease is equally prevalent in northern India, we aimed to assess association to these 26 risk loci in a north Indian population and to see whether dense SNP genotyping in both populations could help refine the association signals. We first performed cross-ethnic mapping in 497 coeliac disease patients and 736 controls from northern India and in 1,150 Dutch cases and 1,173 controls; we used ImmunoChip, a customized genotyping platform providing a higher SNP resolution for immune-related traits than standard GWAS platforms. Due to differences in linkage disequilibrium (LD) structure between Europeans and north Indians, we applied two different strategies which led to replication of 50% of the non-HLA loci. In an 'exact' analysis – directly testing for association of the index SNP – we replicated five loci, whereas in a 'transferability' analysis – testing the exact European SNP along with the variants in LD – we replicated 12 loci. The use of the densely probed ImmunoChip increased the replication rate by 15% compared to a standard GWAS platform. Secondly, we narrowed down the association signals by using a cross-population analysis. We observed that the northern Indian population was characterized by less extensive long-range LD than our Dutch population. For five of the 26 non-HLA loci, this resulted in a finer localization of the association signal and for another five loci we observed a shift in the association signal. It was particularly notable that the long-range LD was broken down at the *IL2/IL21* locus (592 kb) and we were able to localize the association signal to a small interval of 21 kb, near the promoter region of the *KIAA1109* gene.

## Author summary

Most disease-gene mapping approaches have been conducted in populations of European descent. It is unclear how many of these findings can be translated across different racial or ethnic groups: Are the risk alleles the same? Are the effects of associated alleles comparable? How comparable are the localization signals? Fine-mapping is one way to better localize the signals and thereby bring us closer to translational studies and a practical application of genetic findings. Cross-ethnic mapping in distinct populations has the potential to aid such fine-mapping efforts. We applied this strategy to fine-map 26 non-HLA loci associated to coeliac disease in Europeans. For the first time, we used dense genotyping information from a custom-made platform to elucidate the best possible linkage disequilibrium resolution. We replicated 50% of these risk loci in a northern Indian cohort. For five of the regions, the signals from the north Indian population clustered separately from those in Europeans, which allowed us to refine the association signal. In five other loci, we were able to narrow down the association to a smaller genetic interval, covering a cluster of correlated markers. This indicates that we may have identified regions likely to capture the causative variants.

## Introduction

Over the past five years, genome-wide association studies (GWAS) have had tremendous success in mapping genetic loci for common, complex diseases. To date, more than 4,000 SNPs have been reported to be associated with the risk for more than 400 distinct traits (1). Approximately 180 common variants, all with modest effect sizes, have been established for immune-mediated diseases and many more variants are expected to be identified to account for part of the “hidden heritability”. The great majority of GWAS for immune-related diseases have been conducted in populations of European ancestry and it is likely that further risk alleles may be identified in populations with different ethnic backgrounds (2). However, knowledge of the genetic architecture of common complex diseases across multi-ethnic cohorts is rather limited and there is a clear need to replicate disease associations across different ethnicities. A limited number of recent GWAS and replication studies in multi-ethnic cohorts, including African and Asian populations, suggest that Caucasian association signals can be generalized to populations with other ethnic backgrounds (3-5). On the other hand, 25-55% of the

association signals at shared loci are independent between populations, suggesting that disease aetiology is common between populations but that risk variants are often population-specific (6). This observation has consequences for the design of cross-ethnicity replication studies.

Replication studies in ethnically different populations often attempt to replicate only the reported, and therefore most significant, SNP from a certain study or combined meta-analysis. Given the difference in linkage disequilibrium (LD) structure between distinct ethnic groups, such an approach may fail if this top-SNP 'poorly' tags the true risk variant in another population. Alternatively, the most significant and reported SNP, together with SNPs in LD with it, can be tested in additional and ethnically different samples (7). One obstacle is, however, that the available GWAS platforms have been designed to capture genetic variation in populations of Caucasian descent and have less power for other ethnic groups. We aimed to overcome some of these limitations by performing a cross-ethnic study using the most densely probed platform for immune-related diseases available, the so-called ImmunoChip platform.

Coeliac disease is an autoimmune inflammatory disease of the small intestine, caused by consumption of gluten by genetically predisposed individuals. It is the most common form of intestinal inflammatory disorder among Europeans with 1-3% prevalence (8), and a similar 1% prevalence in the northern part of India (9). The largest genetic risk to coeliac disease is conferred by variants in HLA genes, which account for approximately 35-40% of the genetic risk (10). Recent progress in understanding the genetic architecture of coeliac disease has led to the identification of 26 non-HLA loci in a meta-analysis of British, Dutch, Italian and Finnish populations, with replication in further cohorts of European origin (10).

In this study, we focused on the 26 established, non-HLA, coeliac disease loci and assessed their replication in a north Indian population-based sample of 497 patients and 736 unrelated controls. We combined our replication study with fine-mapping, as we used information on 15,851 SNPs across the 26 loci (i.e. on average 610 SNPs per locus) which were present on the ImmunoChip. The ImmunoChip is a custom-made genotyping platform with ~200,000 markers, of which the great majority map within the 183 loci associated to immune-related diseases (11, 12). This platform was specifically designed to fine-map currently known GWAS loci and results with, on average, a 12-15x greater marker density than a standard GWAS chip.

We applied two different strategies to assess the genetic architecture of coeliac disease in the north Indian population: (1) an 'exact' analysis, directly testing the index top-associated SNP reported for Europeans, and (2) a 'transferability' analysis, testing the exact European SNP along with the variants in LD ( $r^2 > 0.05$ ) which were present on the ImmunoChip. Our main objectives were to evaluate the known coeliac disease-associated loci in an ethnically distinct north Indian population, and to refine the association signals. Using these approaches, we convincingly replicated 13 of the 26 loci (50%). Had we tested only the reported European variant using the 'exact' approach, the replication success rate would have been only 19%. We further performed the cross-ethnic fine-mapping by comparing LD structures between the north Indian and Dutch cohorts, followed by haplotype sharing analysis for all 12 transferable loci. We were able to fine map one of the risk haplotypes of the *IL2/IL21* locus, mapping it to a smaller interval near the promoter region of the *KIAA1109* gene.

## Materials and Methods

### Ethical Statement

Ethical approval for this study was granted by the respective institutional and university ethical committees. Informed written consent was acquired from all participants.

### Study Populations

We analyzed two distinct populations: a north Indian and a Dutch population. North Indian patients were recruited from Dayanand Medical College and Hospital, Ludhiana, Punjab, northern India. North Indian controls included blood donors recruited from the same region as the cases and who tested negative for coeliac disease serology. Dutch cases were recruited from University Medical Centre Utrecht, Leiden University Medical Centre, and VU Medical Centre, Amsterdam. A small proportion of samples was recruited via the Dutch patients' society ('*Glutenonderzoek*'). Coeliac disease patients were diagnosed according to standard clinical, serological and histopathological criteria, including a small intestinal biopsy. The majority of samples included in this study have been described in detail elsewhere (12).

## DNA Extraction and Genotyping

The great majority of DNA samples came from blood, while a small proportion of Dutch cases and controls were derived from saliva. Samples were hybridized on the Immunochip, a custom-made Infinium chip with 196,524 markers. Genotyping was carried out according to Illumina's protocol at the genotyping facility, University Medical Centre Groningen. NCBI assembly hg18 was used to map to the genome (Illumina manifest file Immuno\_BeadChip\_11419691\_B.bpm).

## Genotype Data Quality Control

We required samples to have a 98% call rate based on the 172,242 high quality, manually clustered SNPs. Individuals showing high relatedness ( $PI\_HAT > 0.2$ ), or discordant sex were removed. Markers with significant deviation from Hardy-Weinberg equilibrium ( $p > 10^{-3}$ ) or a per-SNP call rate  $< 99\%$  were removed from the final dataset.

Population outliers were identified by multi-dimensional scaling (MDS implemented in PLINK (20)) on 11,192 SNPs that were common ( $MAF > 0.05$ ), LD pruned (a window of 1000 variants, sliding by 10 SNPs at a pair-wise SNP-SNP correlation  $r^2 = 0.05$ ) and shared between Immunochip and HapMap3 samples. MDS analysis was performed jointly with HapMap GIH (Gujarati Indians in Houston, Texas), CEU (Utah residents with Northern and Western European ancestry from the CEPH collection), YRI (Yoruba in Ibadan, Nigeria) and CHB (Han Chinese in Beijing, China) samples to identify major population outliers and also locally. Local analysis was performed separately for Dutch and Indians, to ensure the cases-controls were matching. All outlying samples were excluded from the analysis.

Immunochip contains 3016 variants submitted for the replication of the 'Reading and math skills' GWAS. As these SNPs are unlikely to be confounded by the immunity signal (49 SNPs mapping within HLA were excluded), we used this SNP set as a null reference for calculating genomic inflation.

## Statistical Analysis

Logistic regression was performed using PLINK v1.07 (20). Because we observed population substructure in the north Indian cohort, the north Indian association test was corrected for this by including the first three components of the MDS analysis as covariates in the logistic regression. For both cohorts, gender

was included as a covariate in the logistic regression. 26 non-HLA loci, associated with coeliac disease at genome-wide significance level (10), were the focus of this study. We employed two strategies to test for replication of European signals in Indians.

Exact replication tested for association signals ( $p \leq 0.05$ ) among the 26 index SNPs from Dubois et al. (10) (Table S1). Two of the index SNPs were not present on ImmunoChip and were replaced with their best proxy; within the *IL18RAP* locus, rs917997 was replaced by rs7559479 ( $r^2 = 1$ ) and within the *ITGA4/UBE2E3* locus, rs13010713 by rs1018326 ( $r^2 = 0.9$ ).

Transferability of European signals in Indians was assessed by extracting all correlated markers, measured by  $r^2 > 0.05$  (21) between the index SNP and all variants present within coeliac loci in CEU (from 1000 Genomes data). LD boundaries were defined by extending 0.1 cM to the left and right of the European focal SNP as defined by the HapMap3 recombination map (reported in Dubois et al. 2010 (10)). We then mapped the transferable variants to ImmunoChip and required at least one variant per locus at  $p < 0.01$  to replicate the European signal. The SNP tagging *ICOSLG* locus was not present in the 1000 Genomes data, so we used rs2838531 from HapMap3, a perfect proxy ( $r^2 = 1$ ).

To assess the rate of transferability expected by chance, we permuted case-control labels and mapped the transferable SNPs into the permuted data. For 1000 permutations, 31 loci transferred at  $p < 0.01$ .

Genotype clusters for all the significant markers identified by any of the three approaches were manually inspected in GenomeStudio.

## CROSS Test

Risk haplotypes were established by performing segmental sharing (CROSS test (13)) followed by the association test in PLINK (20). Previously published boundaries of the coeliac disease loci (10) were used and extended by an extra 100 SNPs both upstream and downstream of the locus. We further performed the segmental sharing with the CROSS test, where the longest shared segment between cases versus controls was established after conditioning on each of the SNPs. We considered the results from the CROSS test as significant if the CROSS-p value was lower than the p-value cut-off calculated (separately for each of the loci) as the



Bonferroni corrected value. By conditioning on each of the SNPs in the locus we could investigate its effect on association of the haplotype and select only those SNPs that drove it. Finally, we plotted p-values (single SNP logistic regression test) for each of the SNPs against the CROSS p-value to select candidate segments built up from the SNPs which drive association of the haplotype and which are also strongly associated to the disease (Figure 3 and Supplementary figure 9). Selected in this way, the SNPs are highly likely to include the causative variant or to tag it effectively, creating the risk haplotype. If haplotypes within and between sets show similar sharing conditional on genotypes, it means that there is no evidence for the presence of variation not in strong LD with the SNP's on which conditioning takes place. Simulation tests (not shown here) have shown that when haplotypes span many adjacent loci, phasing genotypes into haplotypes does not improve power. SNPs clustering tightly together were therefore selected for phasing (creating haplotypes) and haplotypes were tested for association using PLINK (20) in each population. To investigate whether the risk haplotypes in the Indian and Dutch populations cluster together, we used the maximum likelihood method implemented in MEGA 5 (22).

### **GWAS and 1000 Genomes Data**

We used previously published coeliac disease GWAS data (10) (Illumina Human Hap550 platform) and extracted 1,284 SNPs mapping in the 26 coeliac loci and shared with ImmunoChip. To assess the transferability rate we followed the same analysis flow as for the ImmunoChip data.

To estimate the coverage of the genetic variation within coeliac disease loci we used 1000 Genomes data (23) (May 2011, SNP calling) and extracted all the annotated variants present in the sequenced samples of European descent: CEU (CEPH individuals), TSI (Tuscan individuals, Italy), FIN (HapMap Finnish individuals from Finland) and GBR (British individuals from England and Scotland). The May release of 1000 Genomes did not include any samples of Indian origin.

### **Test of Heterogeneity**

The homogeneity of ORs between Dutch and Indian cohorts was assessed by the Breslow-Day test. To estimate an accurate significance threshold for this test we permuted the disease status labels in both cohorts and calculated the 5%

quantile of the nominal p-value distribution of Breslow-Day test statistics for 15,851 polymorphic variants in 26 coeliac loci among 1000 permutations.

### Comparative LD Background Evaluation

For 11 transferable loci we visualized the LD structure for the entire region in Indian and Dutch controls using Haploview v4.2 (24). No minor allele pruning was performed to get an accurate account of the cross-population LD differences. In all but three loci, we used all 736 north Indian controls and 1,150 Dutch controls. Due to computational limitations for the *CCR1/5*, *IL2-IL21* and *ITGA4/UBE2E3* loci, we used 400 random controls from each population.

### Estimation of Fixation Index ( $F_{st}$ )

A pair-wise fixation index (25) was calculated for all 12 European index SNPs. We used a control dataset from Indian, Netherlands, UK, Polish, Italian and Spanish populations. We calculated the significance level at  $F_{st} = 0.079$  based on the upper 5th percentile of  $F_{st}$  distribution for 8,007 'neutral' SNPs. We determined neutral SNPs as intronic and uncorrelated with the coding variants (pair-wise  $r^2 = 0.05$ ).

### Software and Resources

Raw genotype data was processed with GenomeStudio\_v2010.3. Statistical analysis and quality control was performed with PLINK v1.07 (<http://pngu.mgh.harvard.edu/~purcell/plink/>). Plots were generated in R (<http://www.r-project.org/>). Haploview v4.2 was used to visualize LD plots. 1000 Genomes data was retrieved from <http://www.1000genomes.org/>. World-wide allele frequency distribution data was taken from (<http://hgdp.uchicago.edu/>).

### Results

We obtained high quality genotype data for 160,448 polymorphic variants that were common in 497 north Indian coeliac disease cases and 736 controls, and in 1,150 Dutch coeliac disease cases and 1,173 controls. The first three components of multi-dimensional scaling analysis (MDS) showed that the north Indian samples overlapped with the GIH samples from HapMap3 (Gujarati Indians in Houston, Texas) (Supplementary figure 1). However, the analysis of GIH and our north Indian samples indicated subtle population substructures. The second component

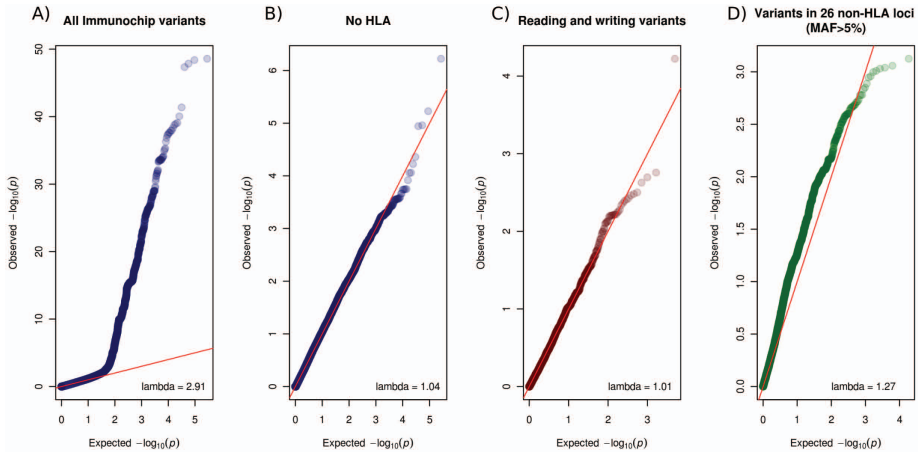
separated the north Indian samples from the GIH samples, while the third component identified further gradients in both these populations. The fact that the first and third components seemed to extract similar information from both the GIH and our cohort indicates that the bias is not due to our cohort sampling but is a genetic characteristic of this ethnic population.

After stringent quality control (see Material and Methods), we obtained 17,979 SNPs mapping within the 26 non-HLA loci, 15,851 of which were polymorphic in at least the north Indian or the Dutch cohort and therefore informative for further analysis. Of the set of 15,851 SNPs, 8,664 (54.7%) were variants with a low minor allele frequency (MAF < 0.05) in the north Indians and 8,484 (53.5%) in the Dutch.

### Replicating European Signals in North Indian Cohort

As expected, we observed strong association of the north Indian coeliac disease cases to the HLA region (rs2854275,  $p = 2.634^{-49}$ ). We observed strong inflation in the test statistics among all ImmunoChip variants ( $\lambda = 2.91$ ), which is mainly driven by the HLA locus (Figure 1A) as exclusion of the HLA region SNPs reduced the  $\lambda$  to 1.04 (Figure 1B). The ImmunoChip is preselected for “immune” SNPs, therefore it is difficult to distinguish enrichment of true signals from population stratification. When we used the 2,626 ‘null variants’ present on the ImmunoChip as a replication of ‘reading and writing skills’ GWAS (see Materials and Methods), we noted that there was no inflation ( $\lambda = 1.01$ ), implying that the observed associations were not confounded by population substructure (Figure 1C). We observed a clear indication of association for the 15,851 SNPs residing in the 26 non-HLA coeliac disease loci ( $\lambda = 1.27$ , Figure 1D).

The initial association of coeliac disease to the 26 non-HLA loci, all of which have modest effects on the genetic risk (OR (odds ratio) of 0.74 -1.36), was performed in a sample size of over 5,000 coeliac cases and 10,000 controls (10). We realize that our north Indian cohort was underpowered to detect associations with small effects although we did reach 80% power for common SNPs (MAF > 0.25) and OR > 1.3 (Supplementary figure 2).



**Figure 1. Inflation in test statistics for all ImmunoChip variants.** We observed strong inflation in the test statistics for all the ImmunoChip variants when including the HLA region ( $\lambda = 2.91$ ; panel A). This inflation decreased to  $\lambda = 1.04$  after excluding some 10,000 markers from the HLA locus, a strong, well-known genetic risk factor for coeliac disease (panel B). The SNPs submitted for replication of the ‘reading and math skills’ were used as ‘null’ variants not confounded by the immune signal, to test for population stratification in our north Indian cohort ( $\lambda$  of 1.01; panel C). Variants in the 26 non-HLA, coeliac disease loci were strongly enriched for association signals (panel D).

### ‘Exact’ Analysis

For the ‘exact’ analysis we aimed to test the same top-associated SNPs reported for Europeans (the ‘index’ SNPs), as described by Dubois et al., (10), in our north Indian cohort. All of the 26 European index SNPs were polymorphic and frequent ( $MAF > 0.05$ ) in north Indians, although the correlation of the minor allele frequencies for these 26 index SNPs between the north Indians and Dutch was much lower ( $r^2 = 0.35$ ) than the correlation between European cohorts (mean  $r^2 = 0.9$ ) (Supplementary figure 3). Of the 26 index SNPs, five were directly replicated at  $p < 0.05$ , with the same direction as in Europeans (Table S1). The five SNPs tagged the following genes: *IL12A* (rs17810546,  $p = 0.02$ ), *LPP* (rs1464510,  $p = 0.048$ ), *THEMIS/PTPRK* (rs802734,  $p = 0.047$ ), *ZMIZ1* (rs1250552,  $p = 0.04$ ) and *ICOSLG* (rs4819388,  $p = 0.035$ ) loci. Of the remaining 21 Indian variants that did not reach the replication significance threshold, 14 (66%) showed the same directionality as in Europeans, which is more than the 50% chance expected under the null hypothesis ( $p = 0.05$ , binomial probability, we excluded two loci for which OR was 1) (Supplementary figure 4).

### ‘Transferability’ Analysis

We next performed an analysis that accounts for the different LD patterns between Europeans and north Indians. In this we tested all the variants in each of the 26 non-HLA loci that were in LD with the top SNP ( $r^2 > 0.05$ , based on CEU, 1000 Genomes Project). To claim a locus as ‘transferable’, at least one SNP per locus had to be associated at  $p < 0.01$  in our north Indian sample. Twelve of the 26 loci were transferable to Indians, which was significantly more than expected by chance ( $p_{\text{permuted}} = 0.031$ ). These were: *PLEK*, *IL18RAP*, *ITGA4/UBE2E3*, *CCR1/5*, *IL12A*, *IL/IL21*, *THEMIS/PTPRK*, *TNFAIP3*, *TAGAP*, *ZMIZ1*, *ETS1* and *ICOSLG* (Table 1); the *IL12A*, *THEMIS/PTPRK*, *ZMIZ1* and *ICOSLG* loci were also identified by the ‘exact’ analysis. None of the transferable variants demonstrated significant cross-population heterogeneity between north Indians and the Dutch (Breslow-Day  $p_{\text{permuted}} < 1.27 \times 10^{-5}$ ). See Table S2 for details of the non-transferred loci.

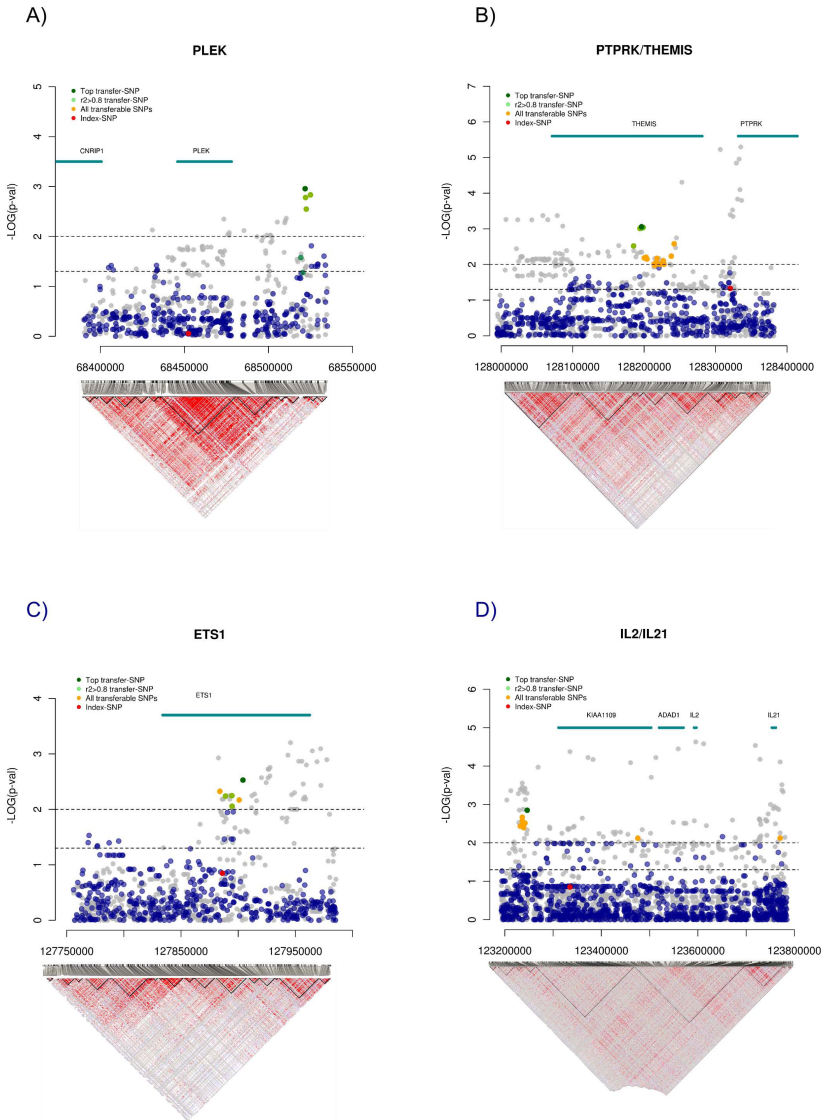
### Coverage and Transferability Rate

To investigate how much we gained by using a dense genotyping platform in contrast to a standard GWAS chip, we calculated the transferability success rate of the 1,284 SNPs that map within the 26 coeliac disease loci that are shared between the Illumina Hap550 platform and ImmunoChip. We replicated nine loci at  $p < 0.01$ , yielding a 35% success rate compared to 50% when using the ImmunoChip (Table S3).

The transferability rate is strongly influenced by the actual coverage of the genetic variation. We therefore assessed how well the ImmunoChip covers total variation at coeliac disease loci in the 1000 Genomes European samples. There are 26,863 polymorphic variants annotated within coeliac disease loci in CEU, TSI, FIN and GBR individuals; 12,497 (46%) of these are shared with ImmunoChip SNPs (821 were non-polymorphic in either the Indians or Dutch). 648 SNPs genotyped on ImmunoChip were not present in the 1000 Genomes data. 5,632 (39%) variants of the non-genotyped 1000 Genomes variants (14,366) contribute to the common variation in Europeans ( $MAF > 0.1$ ). The remaining variation is of low frequency ( $MAF < 0.05$ , Supplementary figure 5), and it is likely that these are largely population-specific or falsely called sequence variants.

**Table 1. Summary statistics of the 12 transferable loci.** Published European SNPs that were not present either on ImmunoChip or in the 1000 Genomes data were replaced by their perfect proxies ( $\delta$ ).

Locus	Index SNP	Top Transferable SNP	Associated allele	BP	MAF		P-value		Odds ratio		$r^2/D'$ with index SNP	
					Dutch	North Indian	Dutch	North Indian	Dutch	North Indian	CEU(1000 Genomes)	North Indian
PLEK	rs17035378 [imm_2_68452459]	rs9309419 [imm_2_68521802]	A	68521802	0.18	0.18	0.043	0.00110	1.16 [1.00-1.34]	0.66 [0.52-0.85]	0.05 0.68	0.03 0.32
IL18RAP	rs7559479 [imm_2_102435219] $\delta$	rs4851005 [imm_2_102377984]	T	102377984	0.39	0.32	0.069	0.00900	0.87 [0.77-0.98]	0.77 [0.64-94]	0.16 1.00	0.33 0.93
ITGA4/ UBE2E3	rs1018326 [imm_2_181716045] $\delta$	NA [imm_2_181849926]	G	181849926	0.36	0.32	0.627	0.00130	0.97 [0.85-1.09]	0.68 [0.54-0.85]	0.21 0.78	0.25 0.62
CCR 1/5	rs13098911 [imm_3_46210205]	rs1799988 [imm_3_46387263]	C	46387263	0.46	0.42	0.041	0.00360	1.12 [1.00-1.26]	0.76 [0.64-0.92]	0.06 0.69	0.03 0.37
IL12A	rs17810546 [imm_3_161147744]	rs1498736 [imm_3_161177252]	A	161177252	0.22	0.39	0.007	0.00293	0.82 [0.72-0.96]	0.75 [0.62-0.91]	0.05 1.00	0.04 1.00
IL2/ IL21	rs13151961 [imm_4_123334952]	rs6534338 [imm_4_123246319]	T	123246319	0.32	0.16	0.004	0.00141	1.20 [1.06-1.36]	1.44 [1.13-1.62]	0.08 1.00	0.03 1.00
THEMIS/ PTPRK	rs802734 [imm_6_128320491]	rs4142030 [imm_6_128196379]	G	128196379	0.38	0.4	0.039	0.00088	1.14 [1.01-1.27]	1.36 [1.24-1.90]	0.22 0.59	0.12 0.47
TNFAIP3	rs2327832 [imm_6_138014761]	NA [imm_6_137980782]	A	137980782	0.49	0.33	0.333	0.00952	0.95 [0.85-1.07]	1.28 [1.06-1.54]	0.08 0.72	0.12 0.59
TAGAP	rs1738074 [imm_6_159385965]	rs9347286 [imm_6_159435348]	T	159435348	0.15	0.08	0.083	0.00167	0.86 [0.73-1.02]	0.55 [0.38-0.80]	0.1 0.79	0.002 0.20
ZMIZ1	rs1250552 [imm_10_80728033]	rs1250549 [imm_10_80730480]	T	80730480	0.47	0.44	0.0005	0.00814	0.80 [0.72-0.90]	1.27 [1.06-1.51]	0.78 0.90	0.75 0.93
ETS1	rs11221332 [imm_11_127886184]	NA [imm_11_127904148]	T	127904148	0.22	0.21	0.242	0.00296	0.90 [0.78-1.04]	0.70 [0.56-0.89]	0.09 1.00	0.08 0.93
ICOSLG	rs2838531 [imm_21_44463014] $\delta$	rs6518350 [imm_21_44446245]	G	44446245	0.19	0.16	0.178	0.00075	0.89 [0.77-1.03]	0.62 [0.47-0.82]	0.48 0.91	0.27 0.67



**Figure 2. Examples of cross-ethnic mapping results.** A) Shifted and more localized association signal in the north Indian samples compared to the Dutch association. Note that the association signal in the smaller Indian samples is stronger than in the Dutch samples, suggesting that some loci may have stronger effects in one population than the other. PLEK is also a locus for which we observed positive selection pressure. B) A cluster of correlated variants localizes in the intronic region of THEMIS gene, whereas the Dutch signal maps 124 kb upstream, pointing towards the 3'UTR of the PTPRK gene. At this locus the index SNP was successfully replicated in north Indians ( $P < 0.05$ ). C) Overlapping signal between the Dutch and north Indians, with tighter SNP clustering over a 21 kb block in north Indians, compared to the dispersed association over a 103 kb block in the Dutch. D) The overlapping signal between the two populations is broadly spread across the region. However, in north Indians a cluster of correlated SNPs localizes in the small 21 kb LD block. Note the LD breakage in north Indians due to the larger number of low frequency SNPs (MAF < 0.1), which results in two smaller LD blocks. Dark blue represents the association signal in the north Indian samples and grey in the Dutch. Dark green depicts the most associated transferable SNP, and light green the SNPs strongly correlated with it ( $r^2 > 0.08$ ). Yellow depicts all the transferable SNPs and red indicates the index SNP. Colours may overlap.

### Cross-Population LD Measure

To assess if the north Indian transferability signals reflect European associations, we estimated the degree of LD correlation (quantified by pair-wise SNP LD,  $r^2$ ) between the European index SNP and the north Indian top transferable SNP in CEU and north Indian data. We observed that the north Indian transferable signals reflected the European LD (correlation coefficient  $r^2 = 0.83$ ) (Supplementary figure 6A). However, the majority of the transferable signals accumulated in the lower tail of  $r^2$  values, with a small shift towards less tight LD in north Indians than in the CEU population.

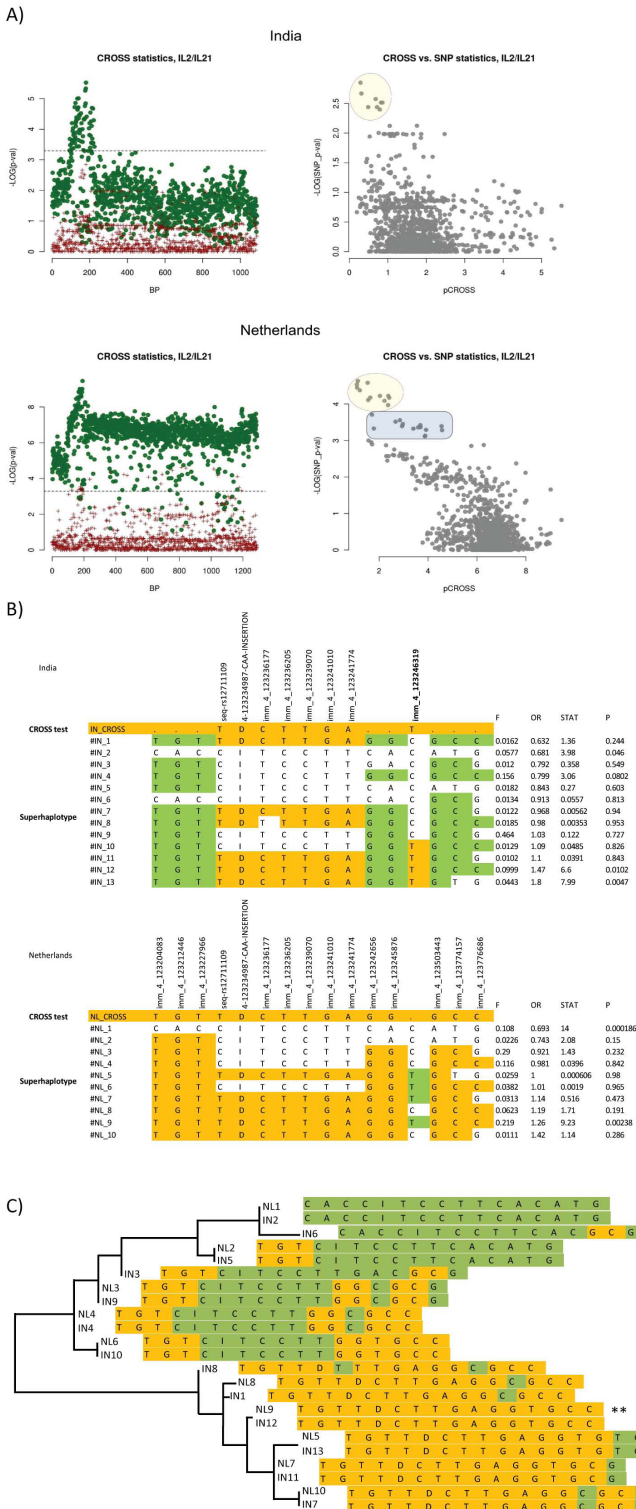
We also evaluated the LD correlation between all transferable variants (all SNPs in  $r^2 > 0.05$  based on CEU, 1000 Genomes data, and present in north Indians) and the top SNP (Supplementary figure 6B). We observed similar results, although the LD correlation between CEU and Indians was lower ( $r^2 = 0.55$ ), and most of the transferable variants were located in the lower tail of  $r^2$  values. We therefore concluded that although north Indians reflect the European association signals, the LD structure at coeliac loci differs substantially between these populations.

### Contrasting Association Signals

As a consequence of longer LD blocks, associations in Europeans often map to regions containing multiple genes and because of the strong LD, it can be difficult to pinpoint the true risk gene. Using ethnically distinct populations offers the opportunity to fine-map signals because of population differences in LD structure.

From the 13 loci that were replicated in north Indians, we excluded the *LPP* locus because it was only replicated by the direct SNP test. We then compared the overlap of the association signal patterns between north Indians and the Dutch and evaluated the LD structure (Supplementary figure 7). Five loci showed a shift in the association pattern (*PLEK*, *ITGA4/UBE2E3*, *THEMIS/PTPRK*, *TAGAP* and *ICOSLG*). At 2p14 (*PLEK* locus) the Dutch signal spreads over the middle LD block and covers the *PLEK* gene, whereas the north Indian signal locates at 69 kb from the European top SNP, downstream of this gene (Figure 2A, Supplementary figure 7). The north Indian signal is located in a block of low LD and is poorly correlated with the European index signal. At the 6q22.33 (*THEMIS/PTPRK* locus) we observed a cluster of correlated variants in the intronic region of the *THEMIS* gene. However, the Dutch signal at this locus maps 124 kb upstream, clearly pointing towards the 3'UTR of the





**Figure 3. CROSS test results at IL2/IL21 locus.**

(A) The left panel compares the association signals of the CROSS test (green dots) and the single SNP associations (red crosses). In Indians, a clearly stronger signal is present in the first part of the locus and identified by both methods. In the Dutch samples, the stronger association in the first part of the region is only observed with the CROSS test, indicating its higher sensitivity. The right panel presents CROSS association results against the single SNP association. SNPs that are highly significantly associated in the single SNP analysis and at the same time close to zero on the pCROSS scale are of interest (meaning the SNP has a strong effect on the haplotype and upon conditioning on it, the association of the haplotype is significantly decreased). SNPs selected by this method were further used for haplotype reconstruction (marked in circles). In the Dutch population we observed two clusters, resulting in two haplotypes. Both scales are logarithmic.

(B) Based on the SNPs selected from CROSS test super-haplotypes were constructed and tested for association. Super-haplotypes were created by merging all SNPs that were selected from both populations. Orange colour represents risk alleles specific for the single population, whereas in green are the risk alleles which were selected from the other population. Only SNPs from the second cluster in the Dutch population were included in the construction of the super-haplotypes (panel A, SNPs in blue circle).

(C) Phylogenetic analysis of the risk types. Orange indicates risk allele and green the protective allele. Stars indicate the two most associated haplotypes in north Indians (IN) and the Dutch (NL).

*PTPRK* gene (Figure 2B; a detailed description of all loci is given in Supplementary figure 7).

At another five loci (*IL18RAP*, *CCR1/5*, *IL12A*, *IL2/IL21* and *ETS1*), the pattern of association in north Indians was consistent with that seen in the Dutch. Nonetheless, for some loci, the association signal could indeed be refined to a smaller region. At the *ETS1* locus, the LD architecture is very similar between the Dutch and north Indians and the cross-population signals overlap (Figure 2C, Supplementary figure 7). However, the north Indian signal is much more tightly clustered in proximity to the top European SNP (21 kb block, from 127882690 bp to 127904148 bp), whereas the Dutch signal is widely spread (over a 103 kb region from 127882690 bp to 127985506 bp). The coeliac disease region at chromosome 4q27 harbours four genes, including two plausible candidates for coeliac disease, *IL2* and *IL21*. Due to the strong LD across the locus, the signal cannot be further refined in Europeans. However, in our north Indian cohort, the LD was broken down due to the larger number of low frequency SNPs (MAF < 0.1) (Figure 2D, Supplementary figure 7). These result in two smaller LD blocks in north Indians, compared with one large block in the Dutch (Supplementary figure 8D). The association signal was also spread along the whole region, although an outstanding cluster of most strongly associated SNPs was located in the small 21 kb block (123246379 bp - 123267309 bp) adjacent to the large block that covers the *KIAA1109* gene and the top European SNP.

At the *TNFAIP3* and *ZMIZ1* loci, only a single SNP was transferable, hence it was not possible to deduce the association patterns.

### **Cross Ethnic Fine-Mapping using CROSS Test and Construction of Risk Haplotypes**

In order to fine map loci replicating in north Indians, we performed haplotype sharing analysis using the CROSS (13) test followed by the haplotype association test. The CROSS test is more sensitive than a single SNP association test as apart from single SNP information, it also incorporates information on haplotype sharing. In short, the CROSS tests if the case and control haplotypes show significantly less sharing than two random haplotypes. The significance for our CROSS analysis was defined by a Bonferroni correction based on LD pruned SNPs ( $r^2 < 0.05$ ) for each locus separately. Three of 13 loci in the CROSS test (we focused on the loci replicating in north Indians) reached the significance threshold in both north Indians and Dutch population: *IL2/IL21*, *IL12A* and *THEMIS/PTPRK* (Figures 3 and S9).

The conditional CROSS test (available from NBIC in an integrated package) compares, using a phenotype randomization test, the agreement between two datasets with the agreement within the two datasets, measured as the number of possible common haplotypes. Possible common haplotypes are defined as the number of markers between the two closest loci with opposite homozygosity. Phasing is therefore eliminated. The comparison is made conditional on the genotypes at a specific locus. If the within set agreement is equal to the between set agreement, the possible shared haplotypes are equal. This test is independent of association. Ideally the most significantly associated SNP should show no difference in the conditional CROSS test, but adjacent linked loci should. If there are multiple causal SNPs, no zero CROSS result is expected. A zero CROSS result coinciding with maximum single SNP association is therefore likely caused by a causal SNP or a SNP in strong LD with a causal SNP (Figures 3 and S9). Further, based on this SNP selection, we constructed haplotypes which were then tested for association (referred to as CROSS haplotypes in Table S5).

*IL2/IL21* was the most successfully fine-mapped locus. It is characterised by a very strong LD in European populations, resulting in the SNP association signal spreading throughout the whole locus in the Dutch population. However, in the north Indians this association is limited to only a small 21 kb region and the pattern of SNP association is mirrored by the CROSS association. This is not the case in the Dutch population where the SNP association covers the whole locus, but the CROSS test is more sensitive and depicts stronger signal overlapping with the north Indians. In the north Indian population, the only risk haplotype (Indian\_CROSS: OR = 1.53,  $p = 0.000562$ ) mapped to the first part of the locus (123236177 bp to 123246319 bp) (Figure 3, Table S5). In the Dutch population we observed two haplotypes, reflected by two clusters of SNPs when we plotted the CROSS results versus single SNP association (Figure 3). The first cluster of SNPs constructed a risk haplotype (OR = 1.36,  $p = 0.000202$ ) comprising 12 SNPs scattered through the entire locus. The second cluster (OR = 1.27,  $p = 0.000432$ ) was made up of 15 SNPs, seven of which overlapped with those in the north Indians, indicating it was the same risk haplotype. Of 15 SNPs, 12 clustered tightly in the small, 41 kb interval (123204083 bp – 123245876 bp). However, three SNPs were located downstream from this cluster, scattering over the locus and reflecting the high LD structure in the region. We constructed a super-haplotype, which comprised SNPs

present in the Dutch and north Indians, and resulted in 16 SNPs (Table S5). The super-haplotype was then separately tested for association in both populations. The most associated risk super-haplotype, #Dutch\_9 (OR = 1.26,  $p = 0.00238$ ), was a comparable risk to the Dutch\_CROSS haplotype (OR = 1.27,  $p = 0.000432$ ). In the north Indians the most associated super-haplotype, #India\_13, showed a lesser degree of association but a higher risk compared to the CROSS haplotype (OR = 1.8 and  $p = 0.0047$  compared to OR = 1.53,  $p = 0.000562$ , respectively) (Table S5). Both the identified risk haplotypes, #Indian\_13 and #Dutch\_9, included the risk allele of the top transferable north Indian SNP rs6534338\*T (imm\_4\_123246319\*T). The CROSS test therefore confirmed the localization of the association signal upstream of *KIAA1109*, which had been identified by the transferability approach. In the phylogenetic analysis, #India\_13 and #Dutch\_9 clustered in the neighbourhood, indicating the same common ancestor.

### Evaluation of the Selection Pressure

To understand evolutionary variability at the 26 coeliac disease loci, we estimated a pair-wise fixation index ( $F_{st}$ ) using the north Indian and Dutch cohorts. We tested  $F_{st}$  for all 26 reported index SNPs (Table S4). We found suggestive evidence of positive selection pressure for four previously reported loci: *IL18RAP* (rs7559479), *CCR1/5* (rs13098911), *PLEK* (rs17035378) and *SH2B3* (rs653178) ( $F_{st} > 0.079$ ).

### Discussion

Recent cross-ethnic studies indicate that a large proportion of disease- or trait-associated common variants established in populations of European descent also contribute to the risk in other ethnic group (3, 4, 14). Yet this observation cannot be generalized because only a few of the GWAS were conducted in non-European populations (15). We sought to replicate and narrow down coeliac disease loci associated in Europeans in a north Indian cohort. We used the Immuchip platform to capture the differences in genetic background between two ethnically distinct populations efficiently. This is the genotyping platform that, at present, offers the highest coverage of regions associated to immune-related disorders. We directly tested for association of the top European SNP (exact test) and also performed association of the 'index' SNP together with variants in LD with it ('transferability' test). Of the 26 coeliac disease loci established in Europeans, five were replicated by

the exact index SNP association analysis and 12 were replicated by the transferability approach, yielding a total of 13 unique, replicated loci. Overall, our replication success rate was 50%, compared to the 19% that we would have obtained by testing only the European index SNP. When investigating the LD correlation between index SNPs and transferable markers, we noticed a trend towards a lower range of  $r^2$  values in north Indians than in the Dutch (Supplementary figure 6). This indicates a lesser degree of tagging properties between the index SNP and transferable markers in north Indians, which could have negatively influenced our replication success rate.

We observed an advantage of using the ImmunoChip over a standard GWAS platform, with a 50% versus 35% successful transferability rate. Nonetheless, despite the high marker density, we estimate that the ImmunoChip only covers approximately 50% of the European genetic variants (based on the 1000 Genomes data, release May 2011) and mainly misses the low frequency variants. Low and rare frequency variants are more likely to be population-specific (15), therefore even if they were included on the ImmunoChip they might be of only limited value. At the moment there is no sequence data available for north Indian populations and it is difficult to estimate how well the ImmunoChip performs in this population. Our transferability success rate is probably underestimated, while differences in haplotype frequencies and LD mean it is likely that the ImmunoChip covers the genetic variation in north Indians less well than in Europeans.

India has been under-represented in genome-wide association studies and we are the first to report an association study on coeliac disease in north Indians and to establish the association of 13 loci in this population. The lack of replication of the remaining 13 European loci may be due to the limited power of our study, and/or poor tagging properties of the causal variant of the tested SNPs in north Indians. However, it is also likely that some European loci will not confer risk for coeliac disease in north Indians. Despite the fact that our north Indian cohort was half the size of the Dutch cohort, the association signals at the *ITGA4/UBE2E3*, *PLEK*, *TAGAP* and *ICOSLG* loci were equal or stronger in the north Indians than those in the Dutch. This indicates that some regions may confer different risk burdens in different populations.

North Indians, although genetically close to Europeans, differ markedly in their allele frequencies from Europeans, which is partly reflected by a lesser degree of LD structure compared to Europeans (16). The different allele frequencies and lower extent of long-range LD in north Indians resulted in different association

signal localization, or allowed finer-scale mapping at some loci. For example, the *IL2/IL21* locus consists of a large LD block of 315 kb that covers the majority of the 591 kb region with highly correlated markers. It contains four genes: two are interleukins, *IL2* and *IL21*, one plays a role in spermatogenesis (*ADAD1*), and one is a transcript of unknown function (*KIAA1109*). Because of its very strong LD, this region has not been fine-mapped successfully using European samples. Our study shows that north Indians have a higher proportion of low frequency markers (MAF < 0.1) than the Dutch, which breaks down the LD structure. The association signal localizes in the small 21 kb LD block (from 123246319 bp to 123267319 bp) adjacent to the larger block covering the *KIAA1109* gene and the top European associated SNP. A recent study (12) showed that there are two independent signals at this locus, with SNPs in LD with the top independent SNPs scattered throughout the region. We noted that neither of the two independent signals were mapping within the identified north Indian risk haplotypes. Although we narrowed down the association signal to the 21 kb block, it does not include a gene, which could suggest that the causal variant plays a regulatory role, and although the associated LD block is in the immediate proximity of *KIAA1109*, the variant may also impact the more distal interleukin genes. This possibility needs to be followed up, preferably by sequencing the region in the north Indians to identify all variants in LD with the top transferable SNP. Thus, although we were not able to pinpoint the likely causal gene in this region, we did succeed in narrowing down the association to a small genetic interval. Another successful example is the *ETS1* locus, where the north Indian signal is limited to tightly clustered variants, suggesting that the causative variant resides within a small 21 kb genetic interval (from 127882690 bp to 127904148 bp).

Differently localized association signals could also indicate locus heterogeneity and different biological pathways underlying the disease aetiology. In our previous GWAS (10), in which we established an association at the 6q22.33 locus, we suggested *THEMIS* as a plausible candidate gene. However, fine-mapping in the Europeans using ImmunoChip clearly points towards the neighbouring *PTPRK* gene, a protein tyrosine phosphatase, whereas the north Indian signal points towards *THEMIS*. Unless functional follow-up studies are conducted to confirm a gene's causality, it will remain unclear which gene truly plays a role in the disease pathogenesis. Or, perhaps, a more complex mechanism including both genes may act at this locus. *PTPRK* and *THEMIS* are both attractive candidates to act in coeliac

disease pathogenesis: *PTPRK*, as a signalling molecule regulating a variety of cellular processes, including cell growth or differentiation, and the neighbouring *THEMIS* gene, which regulates positive and negative T-cell selection during thymocyte development. *THEMIS* is transcribed from the reverse strand and, in fact, both associations could still point to the same gene, for example, in the Europeans by affecting the transcription regulation and in the north Indians by affecting a regulatory mechanism within the gene or by altering the gene structure. This requires follow-up studies for further elucidation.

We found suggestive evidence of positive selection pressure acting on four of the 26 coeliac disease loci: *IL18RAP* (rs7559479), *CCR1/5* (rs13098911), *SH2B3* (rs653178) and *PLEK* (rs17035378) (Table S4) (17-19). We found the strongest selection signal ( $F_{st} = 0.251$ ) at the *SH2B3* locus, which was previously suggested to play a role in bacterial infection (17). The *SH2B3* locus does not replicate in our north Indian cohort, possibly because selection pressure led to a significant decrease in the risk allele frequency of the index SNP (rs653178) in north Indians (11% versus 47% in the Dutch).

We have replicated coeliac disease associations in a north Indian population and also localized and fine-mapped the association signals. However, signatures of positive selection pressure acting on 15% of the loci and our partial replication success could indicate that there are different mechanisms of coeliac disease pathogenesis in the two populations. The modest size of our cohort requires our results to be followed-up in larger samples. We recommend that in any future replication studies in north Indians or other non-Caucasian populations, both the European index SNPs, as well as the top variants from this study, should be tested to better distinguish the relevant signals. Furthermore, our studies should benefit greatly from the 1000 Genomes sequence information for north Indians. This data will allow us to collate information on all the variants that are in LD with the transferable markers identified in our study and to describe more precisely the LD structure and genetic boundaries of variants strongly correlated with associated SNPs. The key to effective fine-mapping is to perform dense genotyping that includes all the known markers present in the north Indians. The results could have an immediate impact by allowing us to refine the associated coeliac disease regions to precise genetic intervals which include the causal variants.

## Acknowledgements

We thank the Dutch clinicians for recruiting coeliac disease patients to give blood samples, as described in our previous studies (C.J. Mulder, G.J. Tack, W.H.M. Verbeek, R.H.J. Houwen, J.J. Schweizer). We thank the UMCG genotyping facility and M. Platteel, K. Fransen and M. Mitrovic for helping generate part of the ImmunoChip data, and S. Jankipersadsing and A. Maatman at UMCG for preparing the samples. We also thank A. Dabral and S. S. Khajuria for DNA isolation and banking at UDSC, New Delhi.

Funding was provided by grants from the Coeliac Disease Consortium, an Innovative Cluster approved by the Netherlands Genomics Initiative, and partially funded by the Dutch Government (BSIK03009 to C.W.) and the Netherlands Organisation for Scientific Research (NWO, VICI grant 918.66.620 to C.W.; Rubicon grant 825.10.002 to A.Z.). S. Senapati is supported by a Senior Research Fellowship from the Council for Scientific and Industrial Research (CSIR), New Delhi, India. We thank Jackie Senior for critically reading the manuscript.

## Reference List

1. Manolio T. A. (2010) Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.*, **363**, 166-176.
2. Bustamante C. D., Burchard E. G., De la Vega F. M. (2011) Genomics for the world. *Nature*, **475**, 163-165.
3. Teslovich T. M., Musunuru K., Smith A. V., Edmondson A. C., Stylianou I. M., Koseki M., Pirruccello J. P., Ripatti S., Chasman D. I., Willer C. J., *et al.* (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, **466**, 707-713.
4. Waters K. M., Stram D. O., Hassanein M. T., Le M. L., Wilkens L. R., Maskarinec G., Monroe K. R., Kolonel L. N., Altshuler D., Henderson B. E., Haiman C. A. (2010) Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups. *PLoS. Genet.*, **6**.
5. Sim X., Ong R. T., Suo C., Tay W. T., Liu J., Ng D. P., Boehnke M., Chia K. S., Wong T. Y., Seielstad M., *et al.* (2011) Transferability of type 2 diabetes implicated loci in multi-ethnic cohorts from Southeast Asia. *PLoS. Genet.*, **7**, e1001363.
6. Fu J., Festen E. A., Wijmenga C. (2011) Multi-ethnic studies in complex traits. *Hum. Mol. Genet.*, **20**, R206-R213.
7. Shriner D., Adeyemo A., Gerry N. P., Herbert A., Chen G., Doumatey A., Huang H., Zhou J., Christman M. F., Rotimi C. N. (2009) Transferability and fine-mapping of genome-wide associated loci for adult height across human populations. *PLoS. One.*, **4**, e8398.
8. Abadie V., Sollid L. M., Barreiro L. B., Jabri B. (2011) Integration of genetic and immunological insights into a model of celiac disease pathogenesis. *Annu. Rev. Immunol.*, **29**, 493-525.
9. Makharia G. K., Verma A. K., Amarchand R., Bhatnagar S., Das P., Goswami A., Bhatia V., Ahuja V., Datta G. S., Anand K. (2011) Prevalence of celiac disease in the northern part of India: a community based study. *J. Gastroenterol. Hepatol.*, **26**, 894-900.
10. Dubois P. C., Trynka G., Franke L., Hunt K. A., Romanos J., Curtotti A., Zhernakova A., Heap G. A., Adany R., Aromaa A., *et al.* (2010) Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.*, **42**, 295-302.
11. Cortes A., Brown M. A. (2011) Promise and pitfalls of the ImmunoChip. *Arthritis Res. Ther.*, **13**, 101.



12. Trynka G., Hunt K. A., Bockett N. A., Romanos J., Mistry V., Szperl A., Bakker S. F., Bardella M. T., Bhaw-Rosun L., Castillejo G., *et al.* (2011) Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.*, **43**, 1193-1201.
13. de Vries A. R., te Meerman G. J. (2010) A haplotype sharing method for determining the relative age of SNP alleles. *Hum. Hered.*, **69**, 52-59.
14. Kurreeman F., Liao K., Chibnik L., Hickey B., Stahl E., Gainer V., Li G., Bry L., Mahan S., Ardlie K., *et al.* (2011) Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records. *Am. J. Hum. Genet.*, **88**, 57-69.
15. Need A. C., Goldstein D. B. (2009) Next generation disparities in human genomics: concerns and remedies. *Trends Genet.*, **25**, 489-494.
16. Reich D., Thangaraj K., Patterson N., Price A. L., Singh L. (2009) Reconstructing Indian population history. *Nature*, **461**, 489-494.
17. Zhernakova A., Elbers C. C., Ferwerda B., Romanos J., Trynka G., Dubois P. C., de Kovel C. G., Franke L., Oosting M., Barisani D., *et al.* (2010) Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection. *Am. J. Hum. Genet.*, **86**, 970-977.
18. Bamshad M. J., Mummidi S., Gonzalez E., Ahuja S. S., Dunn D. M., Watkins W. S., Wooding S., Stone A. C., Jorde L. B., Weiss R. B., Ahuja S. K. (2002) A strong signature of balancing selection in the 5' cis-regulatory region of CCR5. *Proc. Natl. Acad. Sci. U. S. A.*, **99**, 10539-10544.
19. Tang K., Thornton K. R., Stoneking M. (2007) A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS. Biol.*, **5**, e171.
20. Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M. A., Bender D., Maller J., Sklar P., de Bakker P. I., Daly M. J., Sham P. C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559-575.
21. Hoggart C. J., Whittaker J. C., De I. M., Balding D. J. (2008) Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS. Genet.*, **4**, e1000130.
22. Tamura K., Peterson D., Peterson N., Stecher G., Nei M., Kumar S. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.*, **28**, 2731-2739.
23. 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061-1073.
24. Barrett J. C., Fry B., Maller J., Daly M. J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics.*, **21**, 263-265.
25. Weir B. S., Cockerham C. C. (1984) Estimating F-statistics for the analysis of population structure. *Evol.*, **38**, 1358-70.



*Discussion  
and  
future perspectives*

## Where are we now in mapping genetic diseases?

There are around 7,000 monogenic disorders registered in the online Mendelian inheritance in man database (OMIM: <http://omim.org/>) but the causative gene has been identified for around only half of them. Many causal genes have been discovered using positional candidate mapping, i.e. by combining genetic mapping, such as linkage or homozygosity, with the candidate gene approach (1). In the field of complex diseases, huge progress was achieved when the first GWAS (genome-wide association study) was performed in 2005, and since then several hundreds of common loci have been associated to many different complex traits (2). In my thesis I describe the identification of new variants for both monogenic and complex diseases, thereby touching upon a wide range of mapping tools currently in use and showing their limitations and drawbacks. I also touch on the future perspectives of genetics analysis tools and how they could be used in the patient care.

### Linkage analysis and GWAS

Linkage analysis is a powerful tool for mapping candidate regions for well-defined monogenic disorders with a highly penetrant and clear inheritance pattern (3). However, some monogenic diseases can be expressed as complex phenotypes with low penetrance (e.g. familial neuroblastoma (4), or may be seen as sporadic cases due to *de novo* mutations (e.g. recessive autosomal microvillus inclusion disease, chapter 2 in this thesis); applying linkage analysis might then be less than ideal. Moreover, the statistical tools applied in linkage analysis require the effective dataset to have sufficient power, so that one large affected family or many smaller affected families pooled together are needed; locus heterogeneity can be a problem in the latter case if each family is linked to a different genomic location for disease risk.

Microvillus inclusion disease (MIVD) is an autosomal recessive disorder that is presented in chapter 2. Although of high penetrance, its genetic analysis suffers from some of the above bottlenecks, including *de novo* mutations, allele heterogeneity, and a small number of families – since the disease is very rare and fewer than 30 cases have been described in the literature (OMIM: <http://omim.org/>). Fortunately, we could map the gene for MIVD using homozygosity mapping, as one of the families was consanguineous. We found that the *MYO5b* gene was present in one of four large, homozygous regions (> 20 cM) and it became a very strong candidate since it harbored different pathogenic mutations in different

families. More evidence that this gene is the causative one was presented by another research group, who made the same discovery using this approach (5).

Linkage analysis is also not the most appropriate tool for mapping complex diseases as these are highly heterogeneous, of low penetrance, and the causative variants are of rather moderate effect size (6, 7). To overcome this problem, model-free linkage analysis is often used in families segregating for complex disorders. This approach was used in work on celiac disease, (“Exome sequencing in a family segregating for celiac disease”, chapter 4, this thesis). Candidate regions with non-parametrical linkage were established. Although the family was fairly large (four generations), the linkage analysis could still suffer from non-Mendelian inheritance as the statistics for both analyses were not highly significant (non-parametrical linkage values (NPL) < 3). We therefore searched for potential mutations not only in the linkage regions but on an exome-wide scale.

Impressive progress in the genetics of complex diseases has been achieved since 2005 when the first GWA study was performed. Many replication studies followed and some of the GWAS results proved to be replicable in other populations while some were not, indicating heterogeneity plays a role in complex genetic diseases. In chapter 6, “Cross-ethnic replication and fine-mapping of celiac disease loci in a north Indian population”, we described a “locus-wide replication” (called transferability) in an Indian cohort of 26 independent GWAS signals for celiac disease mapped in Europeans. Due to differences in the allele frequencies of variants, as a result of differences in linkage disequilibrium (LD) among the two populations, the direct replication of single variants might not succeed (8). We therefore used *all* the single nucleotide polymorphisms (SNPs) from the 26 loci to investigate their association to celiac disease in the Indian population. In this way we managed to replicate around 50% of the findings in Europeans. GWAS for autoimmune diseases such as inflammatory bowel disorder (IBD) and celiac disease have mapped around 100 common loci, with the majority of these being involved in immune-related pathways (9). The same loci have been associated to many autoimmune disorders, making it clear that there is a common genetic background for such disorders and that the final phenotype is shaped by environmental factors as well as private variants (both common and rare ones) (10). Following on from this work, cross-replication studies were performed where a gene associated to one disease was investigated for another similar disorder. In chapter 5 I described

the cross-replication of a functional variant in the *IL12b* gene for ulcerative colitis. This was a direct replication of a known functional variant that had previously been correlated with other autoimmune (complex) diseases.

Overall, we are not able to map monogenic mutations of more complex inheritance by using candidate positional studies, although for complex genetic diseases, we were able to correlate an impressive number of loci to a disease. However, these loci still explain only a small part of the heritability because they only confer a low risk. In conclusion, monogenic as well as complex diseases are still facing difficult times and new technologies are required to make the discovery rate of causative genes more effective. One of the rapidly evolving, high-throughput tools is next-generation sequencing (NGS), which offers the possibility of discovering a variety of risk variants for both monogenic and complex diseases.

### **Next-generation sequencing: tool for mapping monogenic and complex genetics**

NGS is a high-throughput technique which enables researchers to perform a parallel, fast and effective sequencing of the genome of any individual (11). This method is not biased to any restricted pool of variations and by appropriate data analysis we are able to investigate all the variants together: rare, moderate and common ones (12). NGS enables us to skip the traditional “mapping step” by going directly to interpreting the candidate variants (13), allowing for unbiased investigation of variants regardless of the gene’s function. This is done by applying different prioritization tools in order to distinguish natural, non-pathogenic variants from mutations in the candidate region or genome-wide (14).

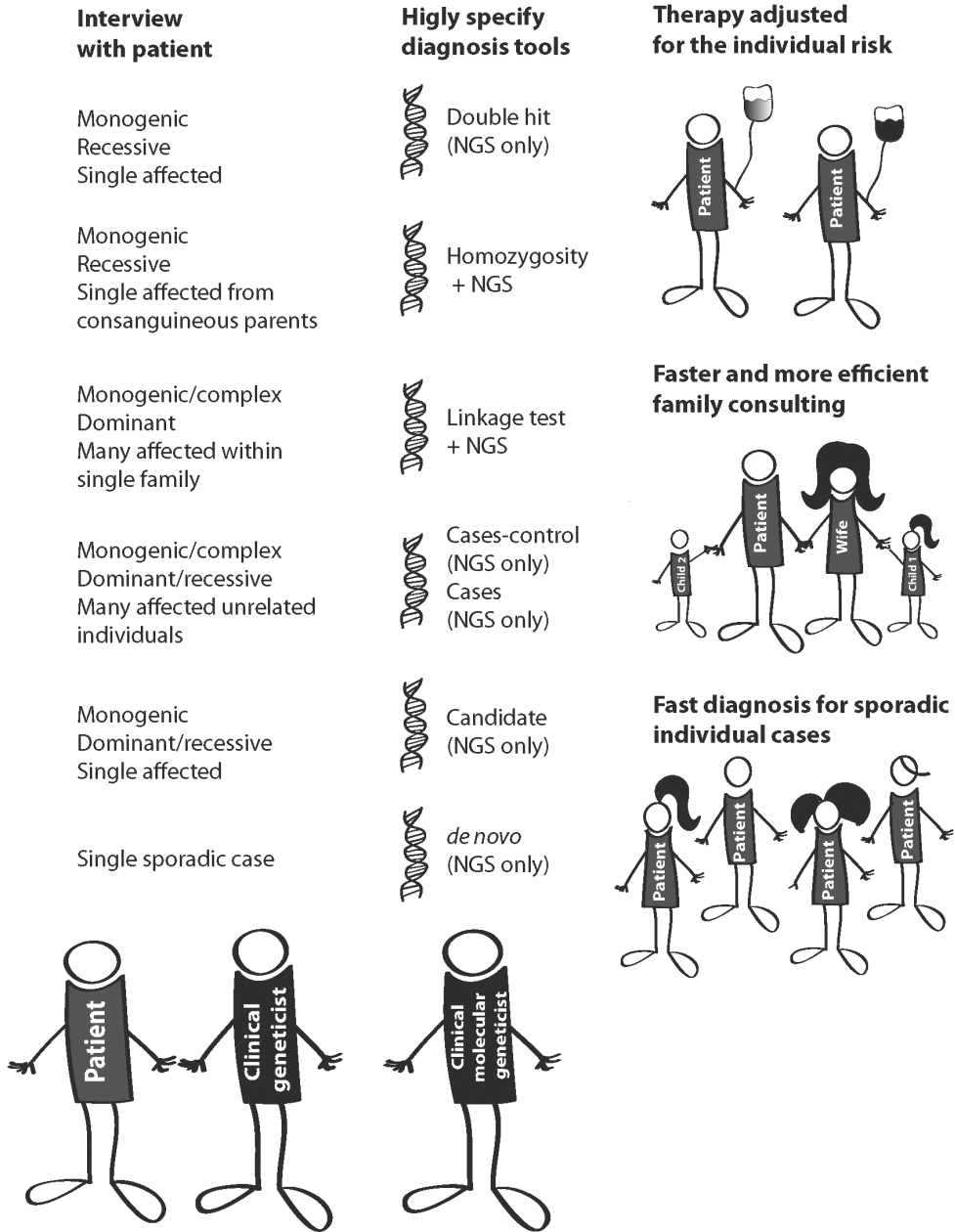
For each type of disease (monogenic or complex), the most important step is to define the character of variants to be prioritized. The easiest to identify are homozygous mutations of very low frequency, thus monogenic recessive diseases are the easiest to analyze. The so-called “double-hit” approach requires only a single individual with a recessive disease and carrying a homozygous, damaging mutation that is not present (or with very low minor allele frequency (MAF)) in ethnically matched controls or public datasets, such as the 1000 Genomes project (Figure 1 and table 1). This approach was used to map rare mutations associated to familial hypolipidemia (15). By applying a double-hit approach to a consanguineous family in our *MVID* study (“Functional characterization of mutations in the myosin gene associated with microvillus inclusion disease”, chapter 2, this thesis) we could

**Table 1 Methods that could be used in the future to diagnose monogenic and complex diseases, depending on the hypothesis adopted and showing the advantages and disadvantages.**

Hypothesis	Method	Advantages	Disadvantages
Homozygous mutation or compound heterozygous	Double hit (NGS only)	One patient, one experiment	Private set of ethnically matched controls
Homozygous mutation is in homozygosity cassette	Homozygosity + NGS	One patient, one experiment	Homozygosity mapping required
Mutation is of high penetrance	Linkage + NGS	No selection bias, 2nd patient sequenced to reduce number of variants	Linkage mapping is required and co-segregation of many variants
Multiple mutation(s) in the same locus (no heterogeneity)	Overlap (NGS only)	No family needed	Heterogeneity (if present). In case-control design need for large ethnically matched sample size
Mutation in gene which function is easily correlated to observed phenotype	Candidate + NGS	No family needed	Biased approach, other additional data needed e.g. gene expression
Mutation present only in offspring	<i>De novo</i> (NGS only)	Detectable	Depends very much on the quality of the data

have skipped the homozygosity mapping and performed only one experiment. In the case of consanguineous families, both methods should be successful, however NGS means such a study is more efficient and simply faster.

In the case of dominant and compound heterozygous mutations, the interpretation of NGS results can be more difficult. This is because a single genome is estimated to carry around 530-610 potentially damaging, loss-of-function variants (in-frame and out-of-frame indels, premature stop codons, and damaging splice sites), with the majority being heterozygous. About 100 of these variants will completely inactivate genes, whereas the rest will influence gene expression to different degrees (16, 17). In chapter 3, this thesis (“True autosomal dominant inheritance of FMF caused by a mutation in exon 8 of the *MEFV* gene”) we presented a family segregating for dominant autosomal disease. After applying filtering steps excluding all known, synonymous variants predicted to be benign we were still left with 125 variants for further consideration and only after applying the knowledge about the disease (candidate approach) we were able to narrow down our analysis to one potentially causative variant. In chapter 4, this thesis (“Exome sequencing in a family segregating for celiac disease”), I described a large family segregating for celiac disease in a dominant-like manner. Although we performed exome sequencing,



**Figure 1 Presenting three steps that could lead to personalized therapy.** 1. An interview with the doctor where a family history is recorded and inheritance is established. 2. The clinical molecular geneticist will choose the most appropriate method for diagnosis. 3. Based on the individual genetic risk, the therapy will be adjusted.

we were not able to identify a causative mutation. One of the reasons might be that we used a too stringent filtering as we only focused on nonsense variants and we could have overlooked the mutation since stop variants only make up a small proportion of the loss-of-function variants. This study was a good example to show that by defining filtering criteria, it is possible to (inadvertently) introduce bias. The best way to solve this problem would be to evaluate each type of sequenced variant (SV) separately using statistical estimation. This can be done by using “weighted” methods, where each of the variants is weighted based on its properties (MAF, damaging score, etc.), and p values are given as a measure of the association to the disease. A huge advantage of these methods is that they also correct for the size of the gene, so that large genes with many SVs do not introduce false-positive hits. By using this approach combined with a filtering-based method, Ionita-Lanz *et al.* were able to map known genes for three monogenic diseases using simulated data (18).

For complex diseases, NGS gives the possibility of simultaneously investigating a wide range of rare variations of moderate and strong effect size, which is not possible using GWAS (2, 19). A case-control set-up has the power to detect causative variants, both common and rare, aggregating in the same gene (Figure 1 and table 1). However, as in GWAS, the sample size is crucial and large case-control studies need to be analyzed in order to pinpoint a true signal. In a study by Surolia *et al.* (20) the authors analyzed data from an inefficient sized cohort of 923 cases, suffering from different inflammatory disorders (such as type 1 diabetes, rheumatoid arthritis, ulcerative colitis, and others) and 648 controls. They reported rare variants in the *SIAE* gene as being associated to the autoimmunity. A replication study performed by a different consortium used a much larger case-control group of 66,924 individuals in total and proved Surolia *et al.*'s finding to be a false-positive (21). Many statistical techniques have been described for associating rare variants in a case-control design to the disease under study. The techniques all show a different performance so it is important to choose the most suitable one in order to minimize false-positive associations (22).

Another possible approach to identifying rare or common variants is to perform a single marker test, as shown by Rivas *et al.* (23). However, the single marker test is, by design, not able to pick up the combined effect of variants on a population level or to detect such a concentration of causative variants in cases where the collapsing (burden) test needs to be used. In the work of Nejentsev *et*



*al.* (24), SVs were found to be associated to type 1 diabetes (T1D) with the use of Fisher's exact test, which directly compares the number of filtered SVs in cases versus controls. However, this and similar methods do not correct for the effect size (odds ratio, OR) of the variants or the MAF, which might be crucial when searching for rare, moderate and common variants in one experiment. Many "weighted analyses" have recently been developed in order to implicate different covariates (25, 26). Using the C- $\alpha$  test (27) that investigates the enrichment of variants by taking into account the direction of the effect, Torgerson *et al.* (28) were able to correlate rare variants in four out of nine candidate genes for asthma susceptibility. Overall, the statistical techniques applied to investigate rare variants for complex diseases in NGS data represent a new field of statistical genetics, which still needs to be improved (29). In chapter 6, this thesis ("Cross-ethnic replication and fine-mapping of celiac disease loci in a north Indian population"), we have described the follow-up of a GWAS analysis based on the ImmunoChip. By using a haplotype sharing approach, we were able to narrow down the regions and pinpoint the risk haplotypes in the *IL2-IL21* locus. If I could perform a follow-up study, I would select individuals carrying the risk haplotypes (cases) or the non-risk haplotypes (controls). Then I would use NGS to target the entire *IL2-IL21* locus rather than exome sequencing, followed by one of the weighted analyses described above, so that as many variables as possible would be taken into account and allow investigation of many different variants in one experiment. In chapter 5, this thesis ("Functional polymorphism in *IL12B* promoter site is associated with ulcerative colitis") I describe an investigation into the coding part of *IL12b* for additional functional variants associated to ulcerative colitis (UC). We did this by traditional Sanger sequencing on some 400 chromosomes. As we did not find any additional variants correlated to ulcerative colitis, the next step would be to sequence the entire *IL12b* locus and, by applying weighted analysis, compare a large group of cases versus controls for the enrichment of functional variants, including non-coding variants, in the entire *IL12b* locus.

Regardless of the experimental set-up, the analysis of NGS data for monogenic or complex diseases presents the problem of how to filter the bulk of non-causative SVs out of the datasets. The most powerful selection step is filtering against private or publicly available NGS datasets. By comparing cases to a carefully matched set of controls, it is possible to reduce the amount of data by about 95%, leaving fewer than 500 variants to be prioritized further (13). Analysis of five

European populations by Zaboli *et al.* (30) showed that 63% of SVs from NGS data were common for all cohorts, while 19% were present only in one of five, i.e. were private, specific variants. Moreover, 62% of novel SNPs (i.e. those not reported before) were highly population-specific (30). In this case, using data only from the 1000 Genomes project is not sufficient and other, more population-specific datasets need to be created or consulted. In study presented in chapter 3, this thesis (“True autosomal dominant inheritance of FMF caused by a mutation in exon 8 of the *MEFV* gene”), I mapped a new mutation for a monogenic, autosomal dominant disease in exome data from two related, affected individuals. After excluding all the private and homozygous variants and filtering the data against the dbSNP database and the 1000 Genomes project, I was able to narrow down the number of variants from 2,225 to 159 (shared by both individuals). It is possible that by using data from ethnically better matched controls, I could reduce this number even further. Exome sequencing of 200 unrelated Danish individuals was sufficient to call variants with a MAF > 0.02 with a 5% false-negative rate (31), which implies that more samples are required for variants with even lower MAF. Recently, the genomes of 250 Dutch trios (parents+child) were sequenced in the Genome of the Netherlands project (GoNL; [www.nlgenome.nl](http://www.nlgenome.nl)) with 12x mean coverage and the data is currently being analyzed. This large, population-specific dataset includes approximately 20 trios per province and will be a source for MAF estimations of rare variants that are exclusive for the Dutch population. Moreover, the trio design allows for a highly reliable validation of variants using identical-by-descent analysis.

Although NGS is a powerful tool to identify new genes, the success rate still depends on the enrichment kit for the exome sequencing. Some genomic loci are not well covered by exome enrichments due either to their high guanine-cytosine (GC) content or simply because they are not included in the enrichment method (13, 32). For example, a causative mutation in *SMOC2* gene was found using the traditional positional-candidate approach, but it was missed by whole exome sequencing as the region was not well enriched (33). A group from Nijmegen did not identify the mutations for Kabuki syndrome (13), whereas Ng *et al.*, investigating the same disease, were able to find causative variants in the *MLL2* gene by using a different enrichment kit (34). In the study presented in chapter 4, this thesis (“Exome sequencing in a family segregating for celiac disease”) we used one of the commercially available kits for exome enrichment. We may have missed the

mutation as the gene was not captured by our kit and it is always worth considering this possibility before deciding on between a whole-genome or whole-exome sequencing approach.

Overall, I think that applying NGS to much of my work would shorten the time taken to discover new mutations (“Functional characterization of mutations in the myosin Vb gene associated with microvillus inclusion disease”, chapter 2, this thesis). If we could use NGS in the work presented in chapter 5, this thesis (“Functional polymorphism in *IL12B* promoter site is associated with ulcerative colitis”) we would investigate the entire *IL12b* locus, including all the non-coding sequences, instead of focusing only on exons of the *IL12b* gene. NGS would be also my choice for continuing the fine-mapping of *IL2-IL21* (“Cross-ethnic replication and fine-mapping of celiac disease loci in a north Indian population”, chapter 6, this thesis).

### **Prospective for NGS and genetics mapping: the personal genome**

NGS is a powerful tool and, when properly applied, it can reduce the diagnosis time to around 9 weeks. Although it is still relatively expensive (whole genome sequencing costs around \$15,000, whereas exome sequencing is around \$3,000), the costs are dropping fast and more diagnostic centers are starting to use this technology for patient care (35). In 2008 NIH’s Undiagnosed Disease Program was begun as a diagnostic program for rare genetic diseases and since then 39 diseases have been diagnosed, of which only three could be resolved with NGS (exome sequencing) (35). This number will soon increase and more diseases will be automatically submitted for diagnosis with NGS. In order to use NGS as a general tool for diagnosis, some bottlenecks still need to be resolved, e.g. the storage of the huge amount of data generated is expensive, and general, well-established diagnosis pipelines for different diseases must be set-up (36, 37). In spite of these problems, some studies have already been published demonstrating the effectiveness and huge potential of NGS in diagnosis.

### **Unknown etiology of the disease**

Diagnosing unknown disorders using NGS is gene discovery and overlaps with fundamental research. To solve the cause of the disease we can apply many combinations of tools commonly used in research with NGS (36). The strategy

will depend on the disorder and resources available. In Figure 1, I present the possibilities for various sets of diagnostic tools to discover the etiology of different genetic disorders.

### **Fast and efficient screening**

NGS offers the possibility to sequence many samples in parallel by using bar-coding and pooling, which can make diagnosis faster and cheaper. To pool many patients and still get good coverage, genomic enrichments are restricted to certain genes that have already been implicated in the disease or in diseases with similar phenotypes. In the work of Corrales *et al.*, 40 patients were pooled and sequenced to determine the causative gene (*VWF*) for Von Willebrand disease, with a discovery rate of 27.5% (the remaining 'results' were due to indels and as the group used single read sequencing, they could not easily pinpoint these) (38). They estimated that, for around 10 kb of sequence, they were able to pool 350 individuals in one experiment and still have sufficient coverage to discover mutations.

### **Re-diagnosis of unresolved cases**

By using exome sequencing Leidenroth *et al.* were able to adjust the diagnosis of patients with facioscapulohumeral muscular dystrophy (FSHD) (39). They did not confirm the mutations in the known causative gene for FSHD, but found well-established risk mutations in the *CAPN3* gene, indicating a case of limb-girdle muscular dystrophy type 2A rather than FSHD. A similar case was described by Choi *et al.* where a child from consanguineous marriage was diagnosed with Barrett's syndrome (40). However, exome sequencing identified a homozygous mutation in the *SLC26A3* gene, associated to other autosomal, congenital chloride diarrhea gene disease, but not in any known genes for Barrett's. NGS in nine additional Barrett's syndrome patients, for whom no mutation had been found in known genes, revealed damaging mutations in *SLC26A3* for five of them. These two examples clearly show the power of using a hypothesis-free NGS method to diagnose patients accurately.

### **Pharmacogenomics**

By performing GWAS scientists were able to adjust drug therapies in, for example, cardiovascular and infectious diseases, based on the common variants discovered. NGS also offers the opportunity of investigating rare variants which

might be more specific for the disease and, based on their genotypes, the treatments could be adjusted more precisely (Figure 1). Most attention is being given to cancer and cardiovascular and immune disorders. Re-sequencing studies have already started that will include many individuals in order to find specific biomarkers which will lead to gene-specific therapies. Recently 3,000 individuals with cancer were enrolled in the USA to undergo exome sequencing of different types of tumors to establish cancer-specific biomarkers (41), while in the large-scale Tumor Sequencing Project (TSP), 632 cancer related genes were sequenced in 188 pairs of tumor and normal tissues in order to find new somatic mutations for lung adenocarcinoma (42). Many of the mutations discovered were found in suppressor genes known to be involved in other types of cancers. These and similar studies will, in the future, allow the type of chemotherapy to be determined by the mutated gene.

In conclusion, we are approaching very exciting times in the field of genetics. New technologies will allow for personalized treatment that should lead to better patient care, although there are still several ethical and practical issues that need to be solved (43). In spite of these, we are now entering a new era of personalized genomic medicine.

## Reference List

1. Ng S. B., Nickerson D. A., Bamshad M. J., Shendure J. (2010) Massively parallel sequencing and rare disease. *Hum. Mol. Genet.*, **19**, R119-R124.
2. Eichler E. E., Flint J., Gibson G., Kong A., Leal S. M., Moore J. H., Nadeau J. H. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.*, **11**, 446-450.
3. Botstein D., Risch N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.*, **33 Suppl**, 228-237.
4. Bourdeaut F., Ferrand S., Brugieres L., Hilbert M., Ribeiro A., Lacroix L., Benard J., Combaret V., Michon J., Valteau-Couanet D., et al. (2012) ALK germline mutations in patients with neuroblastoma: a rare and weakly penetrant syndrome. *Eur. J. Hum. Genet.*, **20**, 291-297.
5. Muller T., Hess M. W., Schiefermeier N., Pfaller K., Ebner H. L., Heinz-Erian P., Ponstingl H., Partsch J., Rollinghoff B., Kohler H., et al. (2008) MYO5B mutations cause microvillus inclusion disease and disrupt epithelial cell polarity. *Nat. Genet.*, **40**, 1163-1165.
6. Manolio T. A., Collins F. S., Cox N. J., Goldstein D. B., Hindorff L. A., Hunter D. J., McCarthy M. I., Ramos E. M., Cardon L. R., Chakravarti A., et al. (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747-753.
7. Risch N. J. (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847-856.
8. Shriner D., Adeyemo A., Gerry N. P., Herbert A., Chen G., Doumatey A., Huang H., Zhou J., Christman M. F., Rotimi C. N. (2009) Transferability and fine-mapping of genome-wide associated loci for adult height across human populations. *PLoS. One.*, **4**, e8398.
9. Lander E. S. (2011) Initial impact of the sequencing of the human genome. *Nature*, **470**, 187-197.
10. Zhernakova A., van Diemen C. C., Wijmenga C. (2009) Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nat. Rev. Genet.*, **10**, 43-55.
11. Pareek C. S., Smoczynski R., Tretyn A. (2011) Sequencing technologies and genome sequencing. *J. Appl. Genet.*, **52**, 413-435.
12. Kazma R., Bailey J. N. (2011) Population-based and family-based designs to analyze rare variants in complex diseases. *Genet. Epidemiol.*, **35 Suppl 1**, S41-S47.
13. Gilissen C., Hoischen A., Brunner H. G., Veltman J. A. (2012) Disease gene identification strategies for exome sequencing. *Eur. J. Hum. Genet.*
14. Piro R. M., Di C. F. (2012) Computational approaches to disease-gene prediction: rationale, classification and successes. *FEBS J.*, **279**, 678-696.
15. Musunuru K., Pirruccello J. P., Do R., Peloso G. M., Guiducci C., Sougnez C., Garimella K. V., Fisher S., Abreu J., Barry A. J., et al. (2010) Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. *N. Engl. J. Med.*, **363**, 2220-2227.
16. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061-1073.
17. MacArthur D. G., Balasubramanian S., Frankish A., Huang N., Morris J., Walter K., Jostins L., Habegger L., Pickrell J. K., Montgomery S. B., et al. (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, **335**, 823-828.
18. Ionita-Laza I., Makarov V., Yoon S., Raby B., Buxbaum J., Nicolae D. L., Lin X. (2011) Finding disease variants in Mendelian disorders by using sequence data: methods and applications. *Am. J. Hum. Genet.*, **89**, 701-712.
19. Maher B. (2008) Personal genomes: The case of the missing heritability. *Nature*, **456**, 18-21.
20. Surolia I., Pirnie S. P., Chellappa V., Taylor K. N., Cariappa A., Moya J., Liu H., Bell D. W., Driscoll D. R., Diederichs S., et al. (2010) Functionally defective germline variants of sialic acid acetyltransferase in autoimmunity. *Nature*, **466**, 243-247.
21. Hunt K. A., Smyth D. J., Balschun T., Ban M., Mistry V., Ahmad T., Anand V., Barrett J. C., Bhaw-Rosun L., Bockett N. A., et al. (2012) Rare and functional SIAE variants are not associated with autoimmune disease risk in up to 66,924 individuals of European ancestry. *Nat. Genet.*, **44**, 3-5.
22. Ladouceur M., Dastani Z., Aulchenko Y. S., Greenwood C. M., Richards J. B. (2012) The empirical power of rare variant association methods: results from sanger sequencing in 1,998

- individuals. *PLoS. Genet.*, **8**, e1002496.
23. Rivas M. A., Beaudoin M., Gardet A., Stevens C., Sharma Y., Zhang C. K., Boucher G., Ripke S., Ellinghaus D., Burtt N., *et al.* (2011) Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.*, **43**, 1066-1073.
  24. Nejentsev S., Walker N., Riches D., Egholm M., Todd J. A. (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, **324**, 387-389.
  25. Feng T., Elston R. C., Zhu X. (2011) Detecting rare and common variants for complex traits: sibpair and odds ratio weighted sum statistics (SPWSS, ORWSS). *Genet. Epidemiol.*, **35**, 398-409.
  26. Zawistowski M., Gopalakrishnan S., Ding J., Li Y., Grimm S., Zollner S. (2010) Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. *Am. J. Hum. Genet.*, **87**, 604-617.
  27. Neale B. M., Rivas M. A., Voight B. F., Altshuler D., Devlin B., Orho-Melander M., Kathiresan S., Purcell S. M., Roeder K., Daly M. J. (2011) Testing for an unusual distribution of rare variants. *PLoS. Genet.*, **7**, e1001322.
  28. Torgerson D. G., Capurso D., Mathias R. A., Graves P. E., Hernandez R. D., Beaty T. H., Bleeker E. R., Raby B. A., Meyers D. A., Barnes K. C., *et al.* (2012) Resequencing candidate genes implicates rare variants in asthma susceptibility. *Am. J. Hum. Genet.*, **90**, 273-281.
  29. Sun Y. V., Sung Y. J., Tintle N., Ziegler A. (2011) Identification of genetic association of multiple rare variants using collapsing methods. *Genet. Epidemiol.*, **35 Suppl 1**, S101-S106.
  30. Zabolni G., Ameur A., Igl W., Johansson A., Hayward C., Vitart V., Campbell S., Zgaga L., Polasek O., Schmitz G., *et al.* (2012) Sequencing of high-complexity DNA pools for identification of nucleotide and structural variants in regions associated with complex traits. *Eur. J. Hum. Genet.*, **20**, 77-83.
  31. Li Y., Vinckenbosch N., Tian G., Huerta-Sanchez E., Jiang T., Jiang H., Albrechtsen A., Andersen G., Cao H., Korneliussen T., *et al.* (2010) Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat. Genet.*, **42**, 969-972.
  32. Teer J. K., Mullikin J. C. (2010) Exome sequencing: the sweet spot before whole genomes. *Hum. Mol. Genet.*, **19**, R145-R151.
  33. Bloch-Zupan A., Jamet X., Etard C., Laugel V., Muller J., Geoffroy V., Strauss J. P., Pelletier V., Marion V., Poch O., *et al.* (2011) Homozygosity mapping and candidate prioritization identify mutations, missed by whole-exome sequencing, in SMOC2, causing major dental developmental defects. *Am. J. Hum. Genet.*, **89**, 773-781.
  34. Ng S. B., Bigham A. W., Buckingham K. J., Hannibal M. C., McMillin M. J., Gildersleeve H. I., Beck A. E., Tabor H. K., Cooper G. M., Mefford H. C., *et al.* (2010) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat. Genet.*, **42**, 790-793.
  35. Maxmen A. (2011) Exome sequencing deciphers rare diseases. *Cell*, **144**, 635-637.
  36. Ku C. S., Cooper D. N., Polychronakos C., Naidoo N., Wu M., Soong R. (2012) Exome sequencing: dual role as a discovery and diagnostic tool. *Ann. Neurol.*, **71**, 5-14.
  37. Gonzaga-Jauregui C., Lupski J. R., Gibbs R. A. (2012) Human genome sequencing in health and disease. *Annu. Rev. Med.*, **63**, 35-61.
  38. Corrales I., Catarino S., Ayats J., Arteta D., Altisent C., Parra R., Vidal F. (2012) High-throughput molecular diagnosis of von Willebrand disease by next generation sequencing methods. *Haematologica*.
  39. Leidenroth A., Sorte H. S., Gilfillan G., Ehrlich M., Lyle R., Hewitt J. E. (2012) Diagnosis by sequencing: correction of misdiagnosis from FSHD2 to LGMD2A by whole-exome analysis. *Eur. J. Hum. Genet.*
  40. Choi M., Scholl U. I., Ji W., Liu T., Tikhonova I. R., Zumbo P., Nayir A., Bakkaloglu A., Ozen S., Sanjad S., *et al.* (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 19096-19101.
  41. Mestan K. K., Ilkhanoff L., Mouli S., Lin S. (2011) Genomic sequencing in clinical trials. *J. Transl. Med.*, **9**, 222.
  42. Greulich H. (2010) The genomics of lung adenocarcinoma: opportunities for targeted therapies. *Genes Cancer*, **1**, 1200-1210.
  43. McGuire A. L., Lupski J. R. (2010) Personal genome research : what should the participant be told? *Trends Genet.*, **26**, 199-201.

Ieder van ons is een uniek individu en wij allen verschillen in uiterlijk, persoonlijkheid, of talenten; sommigen van ons worden vaker ziek en sommigen kunnen harder rennen dan anderen. Er zijn twee belangrijke groepen van factoren die leiden tot de ontwikkeling van deze individuele kenmerken: (1) exogene factoren, dat wil zeggen alle milieu- en andere externe factoren, zoals luchtvervuiling, infecties, of het niveau van het onderwijs, en (2) erfelijke factoren, dat wil zeggen alle informatie opgeslagen in ons DNA. Het verschil in betrokkenheid van elk van deze factoren beïnvloedt de aard van de eigenschap (een fenotypisch kenmerk, gerelateerd aan ziekte of niet-ziekte gerelateerde factoren). Bijvoorbeeld het vermogen om de tong op te rollen is een sterk erfelijke eigenschap die dus afhangt van de genetische informatie in je DNA en niet direct door exogene factoren. Dit soort kenmerken worden monogene eigenschappen genoemd, omdat er slechts een gen is die de eigenschap bepaald en de invloed van exogene factoren zeer beperkt is. Een ander voorbeeld is de body mass index (BMI), die wordt bepaald door meerdere genen overgeërfd van je ouders (bijvoorbeeld de genen die de stofwisseling reguleren) en door de omgeving en levensstijl (bijvoorbeeld dieet, lichaamsbeweging). Dit soort eigenschappen worden complexe eigenschappen genoemd, omdat ze beïnvloedt worden door tal van genetische en omgevingsfactoren.

Elke omgeving- of genetische verandering kan invloed hebben op deze eigenschappen en kan leiden tot ziekte. Om ziekten te voorkomen en te genezen heeft de medische wetenschap veel energie gestoken in het ontdekken en het uitleggen van de bijdragen van zowel omgeving- als genetische factoren voor de menselijke gezondheid en ziekten. Het belangrijkste doel van epidemiologische studies is om veranderingen in de omgeving aan verschillende menselijke eigenschappen te correleren, terwijl de medische genetische studies focussen op het onderzoeken van afwijkingen in ons DNA bij verschillende complexe (bv. kanker, diabetes type 1, coeliakie) en monogene aandoeningen (zoals hemofilie uit te leggen, de ziekte van Huntington).

In mijn proefschrift presenteer ik een breed scala aan technieken die vaak gebruikt worden in huidig medisch genetisch onderzoek naar genen en mutaties betrokken bij verschillende ziekten met verschillende genetische achtergronden. In hoofdstuk 1 introduceer ik de twee belangrijkste groepen van genetische aandoeningen: monogene en complexe aandoeningen. Ik presenteer de kenmerken van de genetische componenten die ten grondslag liggen aan beide groepen van



genetische ziekten, de verschillen tussen hen, en toon de betrokkenheid aan van genetische en omgevingsfactoren in de ontwikkeling van elke ziekte. Ik leg ook de meest gebruikte methoden uit om de oorzakelijke genen voor beide groepen van ziekten in kaart te brengen, en ik benoem hun belangrijkste knelpunten.

In de hoofdstukken 2 en 3 presenteer ik twee monogene ziekten en twee verschillende toegepaste technieken om hun causale genen te achterhalen. In hoofdstuk 2 wordt een geval besproken van recessieve microvillus inclusieziekte in een consanguine familie (de ouders van de getroffen kinderen waren eerstegraads neef en nicht). Bij deze ziekte wordt de opname van voedingsstoffen uit de dunne darm bij een pasgeborene ernstig verstoord. De oorzakelijke mutatie in het gen *Myo5b* is in kaart gebracht met behulp van de positionele kandidaatgen aanpak: kennis over eiwit-expressie bij patiënten in combinatie met hypothesevrije genetische mapping stelde ons in staat om dit oorzakelijke gen te vinden. Hoofdstuk 3 beschrijft een dominante monogene ziekte, familiale mediterrane koorts, overervend in een familie over drie generaties. Deze ziekte is een immunologische aandoening die wordt gekenmerkt door periodes van hoge koorts en andere symptomen, zoals huiduitslag. Deze studie toonde de kracht van een vrij nieuwe, hypothesevrije, next-generation sequencing methode, waarbij het gehele exoom van een individu werd gesequenced en stapsgewijze filtering werd toegepast op meer dan 20.000 varianten om een oorzakelijke mutatie te identificeren in exon 8 van het *MEFV* gen.

Hoofdstukken 4, 5 en 6 beschrijven twee complexe ziekten waarbij de dunne darm is aangedaan: coeliakie, waarbij gluten een sterke immuunreactie in de dunne darm veroorzaakt (hoofdstukken 4 en 6) en colitis ulcerosa, een immuun-gerelateerde ziekte van de dunne darm waarbij de omgevingsfactor nog niet bekend is, maar die ligt in een overmatige immuunreactie op de darmflora. Hoofdstuk 4 beschrijft een grote familie (vier generaties) van Nederlandse afkomst waarin coeliakie overerft. De prevalentie van de ziekte is zeer hoog in deze familie en maar liefst 40% van de nakomelingen in de eerste generatie is aangedaan, wat kenmerkend is voor monogene dominante ziekten. We hebben daarom gezocht naar een sterke mutatie in het hele exoom in next-generation sequence data. Hoewel we niet in staat waren om de oorzakelijke mutatie te vinden, was deze studie erg belangrijk omdat het de complexiteit en de moeilijkheden aantoont van het beperken van een vermoedelijke genetische regio om een oorzakelijke variant te identificeren in families waarin een complexe aandoening met een Mendeliaans patroon overerft.

Hoofdstukken 5 en 6 illustreren verschillende benaderingen van replicatie voor het werk beschreven zoals in hoofdstuk 4. In hoofdstuk 5 wordt een “ziekte-kruisende” benadering toegepast, waarbij we de associatie van een functionele variant in de promotor regio van *IL12B* met colitis ulcerosa gerepliceerd hebben. Dezelfde variant was al in verband gebracht met andere immuungerelateerde ziekten.

In hoofdstuk 6 wordt een “eticiteit-kruisende” replicatie studie beschreven waarin varianten geassocieerd met coeliakie in een Nederlands cohort werden onderzocht in een ander cohort van een andere etnische afkomst (Noord-Indiërs). Deze studie toonde de kracht van het gebruik van locus replicatie in plaats van enkele variant replicatie, omdat verschillen in “linkage disequilibrium” tussen populaties foutnegatieve resultaten kunnen introduceren. Met behulp van een haplotype-sharing methode hebben we het associatie gebied van het *IL2-IL21* locus teruggebracht met een factor 30, waardoor fine-mapping van de oorzakelijke variant beter mogelijk is gemaakt.

In hoofdstuk 7 bespreek ik de behaalde resultaten in de humane genetische studies van monogene en complexe ziekte, en presenteer ik een toekomstperspectief op het in kaart brengen van genen met de focus op next-generation sequencing.

Each of us is a unique individual and we all differ in looks, personality, or talents; some of us get sick more often or can run faster than others. There are two main groups of factors that lead to the development of these individual characteristics: (1) exogenous factors, i.e. all environmental and other external factors, such as air pollution, infections, and level of education, and (2) inherited factors, i.e. all the information carried in our DNA. The differential involvement of any of these two types of factors influences the character of the trait (a trait is a phenotypic characteristic, related to disease or non-disease factors), for example, the ability to roll your tongue is a strongly inherited trait and depends on the genetic information in your DNA, so exogenous factors cannot influence it directly. These kinds of traits are called monogenic traits, because only one gene is controlling it and the influence of exogenous factors is very limited. Another example is your body mass index (BMI), which is defined by multiple genes inherited from your parents (e.g. the genes controlling metabolism) and by the environment you live in (e.g. diet, exercise). These kinds of traits are called complex traits as they are influenced by many genetic and environmental factors.

Any environmental or genetic change can affect such traits and may lead to disease. To prevent and cure diseases, medical science is putting much effort into discovering and explaining the contributions from both environmental and genetic factors to human health and diseases. The main goal of epidemiological studies is to correlate changes in environment to different human traits, whereas medical genetic studies focus on investigating aberrations in our DNA to explain different complex (e.g. cancer, type 1 diabetes, celiac disease) and monogenic disorders (e.g. hemophilia, Huntington's disease).

In my thesis I present a wide spectrum of tools commonly used in current medical genetic studies to identify the genes and mutations that lead to diseases governed by different types of genetic architecture.

In chapter 1, I introduce the two main groups of genetic disorders: monogenic and complex disorders. I present the characteristics of genetic components underlying both groups of genetic diseases, the differences between them, and have illustrated the involvement of genetic and environmental factors in the development of each disease. I also explain the most common methods used to map the causative genes for both groups of diseases, and I assess the main bottlenecks affecting these methods.

In chapters 2 and 3, I present two monogenic diseases and two different tools used to discover their causal genes. Chapter 2 presents a case of recessive microvillus inclusion disease in a consanguineous family (the parents of the affected child were first-degree cousins), in which uptake of nutrition from the small intestine in a newborn was disturbed. The causative mutation in the gene *Myo5b* was mapped using the positional-candidate approach: knowledge about protein expression in patients combined with hypothesis-free genetic mapping enabled us to find this causative gene. Chapter 3 presents a dominant monogenic disease, Familial Mediterranean Fever, segregating in a family over three generations. This disease is an immunological disorder characterized by periods of high fever and other symptoms, such as skin rash. This study showed the power of a fairly new, hypothesis-free, next-generation sequencing method, in which the entire exome of an individual was sequenced and step-wise filtering was applied to more than 20,000 variants in order to identify one causative mutation in exon 8 of the *MEFV* gene.

Chapters 4, 5 and 6 present two complex diseases that affect the small intestine: celiac disease, in which gluten triggers a strong immune response in the small intestine (chapters 4 and 6) and ulcerative colitis, an immune-related disease of the small intestine for which the environmental factor is still unknown, but which may lie in an excessive immune response to the intestinal microflora. Chapter 4 describes a large family (four generations) of Dutch origin segregating for celiac disease. The prevalence of the disease is very high and affects as many as 40% of the offspring in the first generation, which is characteristic for monogenic dominant diseases. We therefore looked for a strong mutation in the exome-wide next-generation sequence data. Although we were not able to find the causal mutation, this study was very important since it demonstrated the complexity and difficulties of narrowing down a suspected genetic region to a causative variant in families segregating for complex disorders with a Mendelian-like inheritance pattern.

Chapters 5 and 6 illustrate different replication approaches for the work described in chapter 4. In chapter 5 we applied a cross-disease approach, in which we replicated the association of a functional variant in the promoter site of *IL12b* with ulcerative colitis. The same variant had already been associated with other immune-related diseases.

Chapter 6 describes a cross-ethnic replication study in which variants

associated to celiac disease in a Dutch cohort were examined in another cohort of different ethnic origin (North Indians). Here we showed the power of using locus replication rather than single variant replication, since differences in linkage disequilibrium between populations might introduce false-negative results. Using a haplotype-sharing method, we narrowed down the association region at the *IL2-IL21* locus by a factor of 30 , thereby making fine-mapping to the causal variant more feasible.

In Chapter 7 I discuss the achievements in human genetic studies of monogenic and complex disease, and present a future perspective on gene mapping with much of the focus on next-generation sequencing.

Każdy z nas jest inny. Różnimy się wyglądem, charakterem, talentami. Tak też niektórzy z nas chorują częściej niż inni, a jeszcze inni szybciej biegają, itp. Można wyróżnić dwie główne grupy czynników, które wpływają na każdą cechę. Pierwszą stanowią czynniki egzogenne, czyli środowiskowe np. zanieczyszczenie powietrza, przebyte infekcje czy edukacja. Druga grupa to czynniki dziedziczne, czyli cała informacja zawarta w naszym DNA. Czynniki obu grup mogą mieć różny udział w formowaniu cech, dlatego też cechy dzielimy na: cechy determinowane jednogenowo, czyli silnie dziedziczne, gdzie czynniki środowiskowe nie mają wpływu na ich kształtowanie lub wpływ ten jest bardzo ograniczony (np. zdolność układania języka w rurkę). Z drugiej strony mamy cechy złożone (wieloczynnikowe), gdzie zarówno czynniki środowiskowe, jak i czynniki dziedziczne odgrywają znaczną rolę. Na przykład współczynnik masy ciała (z ang. BMI, body mass index) jest cechą w części dziedziczną (geny odpowiedzialne za metabolizm organizmu), a w części kształtowaną przez czynniki środowiskowe (dieta, ruch i regularne ćwiczenia).

Każda zmiana w obrębie czynników środowiskowych albo dziedzicznych może prowadzić do zmiany cechy i w konsekwencji do choroby i spadku komfortu życia. Aby zapobiegać i leczyć wywołane w ten sposób choroby, niezbędne jest badanie wpływu czynników dziedzicznych i środowiskowych na nasze zdrowie. Wyodrębniamy dwie nauki: epidemiologię, która bada wpływ czynników egzogennych oraz genetykę medyczną badającą wpływ czynników dziedzicznych (zmian w DNA, czyli mutacji) na choroby złożone (np. nowotwory złośliwe, cukrzycę typu I, celiakię) i choroby jednogenowe (np. hemofilię, chorobę Huntingtona).

Niniejsza praca doktorska prezentuje szerokie spektrum metod używanych w genetyce medycznej w celu identyfikacji czynników dziedzicznych, tzw. szukanie genów kandydatów, powodujących choroby jednogenowe i złożone

W rozdziale 1. opisuję dwa typy chorób: złożone i jednogenowe z uwzględnieniem podstawowych różnic, zarówno w czynnikach genetycznych, jak i w ich odmiennym współdziałaniu z czynnikami środowiskowymi. W dalszej części rozdziału 1. ilustruję najczęściej używane techniki służące do poznania czynników dziedzicznych, które powodują daną jednostkę chorobową (mutacje w DNA) oraz przedstawiam najczęstsze przyczyny niepowodzenia metod.

W rozdziałach 2. i 3. przedstawiam dwie choroby jednogenowe i dwie różne metody użyte w celu zidentyfikowania mutacji w genie kandydacie. W rozdziale 2. opisuję wrodzoną, autosomalną, recesywną atrofię mikrokosmków

(z ang. microvillus inclusion disease, MVID) zdiagnozowaną u dziecka pochodzącego z małżeństwa kuzynów pierwszego stopnia. U chorych obserwuje się zaburzone pobieranie pokarmu z jelita cienkiego. Mutacja w genie *Myo5b* została zidentyfikowana metodami przeszukiwania genomu oraz wyznaczaniu genu kandydata (z ang. positional-candidate cloning). Dzięki tym metodom byłam w stanie zidentyfikować region w genomie, gdzie prawdopodobieństwo znalezienia mutacji jest większe, a dodatkowa wiedza na temat ekspresji białek w tkance pacjenta pozwoliła mi na wskazanie jednego genu kandydata w tym regionie. W rozdziale 3. przedstawiłam autosomalną, dominującą chorobę wrodzoną zwaną rodzinną gorączką śródziemnomorską zdiagnozowaną u członków trójgeneracyjnej rodziny. Choroba ta powoduje okresowe napady wysokiej gorączki i inne symptomy np. wysypkę na skórze. Praca ta ukazuje wielkie możliwości związane z nową techniką przeszukiwania genomu ludzkiego zwaną następną generacją sekwencjonowania (z ang. next generation sequencing, NGS). Cała informacja genetyczna kodująca białka została zsekwencjonowana w krótkim czasie i wszystkie zmiany w DNA pacjenta zostały odpowiednio opisane. W kolejnej fazie selekcyjnej jedna z 20,000 opisanych zmian w egzonie 8 *MEFV* genu kandydata została zidentyfikowana jako powodująca chorobę.

W kolejnych rozdziałach opisałam dwie choroby złożone: celiakię, która spowodowana jest silną reakcją autoimmunologiczną w obecności glutenu w jelicie cienkim (rozdziały 4. i 6.) oraz ulcerozę, chorobę immunologiczną jelita cienkiego, dla której czynnik środowiskowy nie został poznany, chociaż może być silnie związany z mikroflorą jelita pacjenta.

W rozdziale 4. prezentuję czteropokoleniową rodzinę pochodzenia holenderskiego, w której celiakia występuje nadzwyczaj często i obejmuje 40% potomstwa z pierwszego pokolenia, co wskazuje na model dziedziczenia podobny do modelu dominującego, autosomalnego, w chorobach jednogenowych. Przyjmując ten model dziedziczenia, zastosowałam szybką i wydajną metodę NGS, gdzie w fazie selekcyjnej szukałam mutacji, która powodowałaby wysokie dziedziczenie choroby. Mimo, że nie udało mi się znaleźć mutacji, moja praca porusza ważną kwestię, ukazując szeroki wachlarz trudności związanych z przeszukiwaniem genomu w rodzinach cierpiących na złożone choroby o charakterze chorób jednogenowych.

W rozdziałach 5. i 6. przedstawiam różne metody używane w celach replikacji poprzednio uzyskanych wyników. W rozdziale 5. opisuję replikację

funkcjonalnej zmiany w *IL12b* genie kandydacie w ulcerozie. Ta sama funkcjonalna zmiana została już poprzednio skorelowana z innymi immunologicznymi chorobami (tzw. replikacja międzychorobowa). Natomiast w rozdziale 6. zaprezentowałam replikację międzypopulacyjną, gdzie zmiany w DNA skorelowane z celiakią w populacji holenderskiej zostały przetestowane w populacji hinduskiej z północy Indii. W pracy tej bardzo wyraźnie pokazałam wyższość replikacji wszystkich zmian w DNA z regionu kandydata, w przeciwieństwie do replikacji jednej, poprzednio najbardziej skorelowanej zmiany. Jest to ważne ze względu na różnice w wielkości regionów kandydatów, gdyż testując jedną zmianę, możemy otrzymać wynik fałszywie negatywny. W dalszej części rozdziału opisałam metodę wykorzystującą różne długości haplotypów w regionie kandydacie w celu jego zawężenia. Używając tej metody, zmniejszyłam region zawierający geny *IL2-IL21* trzydziestokrotnie, dzięki czemu dalsze prace nad wskazaniem genu kandydata w tym regionie mogą być bardziej wydajne.

W rozdziale 7. opisuję i dyskutuję najnowsze osiągnięcia w identyfikacji genów kandydatów dla chorób jednogenowych i złożonych. Ostatecznie skupiam się na perspektywach, silnie odnosząc się do techniki, NGS, która moim zdaniem jest przyszłością genetyki klinicznej.




**Dear all,**

I remember coming to Groningen for the first time for an interview. It was an intensive but pleasant day. I talked to many people from, at that time, a still rather small celiac group, including some technicians, Martin W. and Cisca. In the end, Cisca invited me to join her group for four years. So I became a PhD student in the Department of Genetics. I was very enthusiastic but very surprised to discover that everybody spoke a form of “Chinese”... well, the genetics at that time were very strange for me and I realized very quickly that I knew more or less NOTHING about human genetics and all these strange GWAS and linkage studies. My PhD period went by fast and I learned to speak a bit of the language of genetics. But this period has been much more than just genetics. It was also me learning about people, making friends and, in the end, it was a very intensive lesson about myself. I think I became more independent and more self-confident. That lesson was interesting, sometimes very funny and sometimes sad ... just like I am: full of emotions :-). At this point I would like to say a sincere “thank you” to all the people I met during this period: my supervisors, co-workers, friends and family, as you gave me strength and many beautiful moments. I will keep all the memories about YOU close to my heart.

Dear Cisca and Marten, from the very beginning you helped me. I stayed in your home, nowadays known as “*Hotel Genetica*”. I think the majority of the kitchen stuff I used in my home in Groningen over these past few years came from you. My favorite lamp, I got from you and I brought it back to Poland ... Cisca, thank you for being a very wise supervisor. I could always count on you when it came to science, and also when I had my difficult times. I really consider it so cool that your room is always open to us and although you are VERY busy, you still somehow (I personally think it’s magic) find time to meet up with your PhD students on a regular basis. You taught me how to analyze data, and think and write about science (talking about science is still a problem :-)). I am very glad and thankful I could be your student.

Dear Martin W., I really admire your sense of humor and I very much enjoy talking to you as I think we share many interests: history books, a passion for music, and good food. People think I am talkative ... well I think you beat me :-), I so much enjoyed your enthusiasm for science. You were my supervisor for only a short time, but the project we worked on together is very special for me and I think, in the end, it will influence my future choices. Thank you for your guidance and believing in me.

Dear Cleo, my supervisor, my friend. There are so many things I would like to thank you for, but I have to save the space as printing this thesis is expensive :-). You are really the person who can calm me down. I think we can talk about everything and I much appreciate the honesty between us. I enjoyed being your student and writing the *IL12b* paper, although it was short, it was important because it gave me a huge burst of energy. I admire you and I think you are a very strong person with lots of love and a very good skill for making delicious food. Lots of kisses to your lovely girls.



I am grateful to Prof. A.M.H. Boots, Prof. P.C. Limburg, Prof. A.J.H. Moshage and Prof. E.H.H.M. Rings for agreeing to be part of the reading committee for my thesis, for taking the time to read it, and for giving their approval for the defense.

Dear GUIDE team, Prof. A.J.H. Moshage, Riekje Banus, Maaïke Bansema, Peter G. Braun, Mathilde T.L. Pekelaer, thank you for all your help during these years. Prof. Moshage, I would like to thank you for good and helpful project management course. Dear Riekje, thank you for the honest conversations.

Dearest Jackie, I am so grateful for all the corrections and editing of my English. I very much appreciate you always being ready to help me, even when we did not have an appointment (sorry for that). I wish you and your husband a lot of fine moments, and many beautiful holidays.

Dear “MVID team”, we went through an emotional time during the writing and submitting of the MVID paper. Sven, thank you for collaborating with me, in the end it was a success. Edmond, I appreciate your updates about patients and the great dinner for our MVID team. Magdaleno, jako pierwsza obroniłaś nasze wyniki. Dziękuję za wspólną pracę i życzę Tobie wiele sukcesów w przyszłości.

Dear collaborators from far-away Finland, thank you for having me in your department. Dear Päivi, thank you for the opportunity to work with you. I am so happy we could write a paper together. Dear Elisabeth, you are the person who helped me a lot in the lab and also in my private life. I am so happy we have stayed in touch. Dear Lotte, Katja, Amarjit and Hanne I enjoyed having some very tasty lunches and/or conversations with you.

There are more than 200 people working in the Department of Genetics at the UMCG. Thank you all for a cool department day out and for all the nice gatherings to celebrate Christmas and New Year. Dear Robert, thank you for support during the difficult times. Richard, I would like to thank you for helping me with writing the letter of motivation to apply for the clinical medical geneticist position. Charles, I really appreciate that you pay attention to people and always remember people’s names. Ellen N., it was very nice to have conversations with you over lunch. Dear Brigit, thanks for explaining, over a delicious dinner, what the responsibilities of a clinical medical geneticist are. Dear Conny, thank you for some useful tips before visiting Joris in Belgium. Dear Marina, Edwin and Hayo, I would like to thank you for all your help with the administration, finances and computers. Dear Mentje, Ria and Joke, thank you for all the help you gave me with copying, sending and receiving documents. Bote, I really enjoyed your good spirit and optimism.

Special thanks to you, H el ene, for helping me with my thesis, and the submissions and appointments. Although you have so many things to deal with, you are always helpful and greet everyone with a huge smile. Justyn, dzi ekuj e za wspania e lekcje reggeatonu. Wspania e si  bawi am. Dear Helga and Annemieke, I enjoyed your dry, sarcastic and very intelligent sense of humor, just like mine :-). Dear Marlies, I really


enjoyed our after-work outings for fine food and wine. Dear Angela, I enjoyed our conversations about relationships and your delicious carbonara pasta. Dear Pieter (van der Vlies), I really appreciated your help with using a bolt cutter (*betonschaar*) and getting rid of stuff from my old house. Dear roommates: Aśka, Jihane, Paul, Dineke, Javier, Mats, Duco, Olga, Suzanne, Céline, and Cleo, thank you for a pleasant time and all the fine conversations. Dear Paul, I enjoyed talking to you, I really like your sense of humor. I wish you a lot of success in your future career. Duco, it was cool to hang out, drink beer, and eat sushi together. Dear Jantine, Michiel, I enjoyed talking to you over a beer. Rutger, you are a funny, good looking guy, thanks for that :-). It was a pleasure seeing you maturing ;). Karen T., I appreciated your “straight-to-the-point” comments and good humor. I wish you all the best with opening a guesthouse.

Dear friends, co-workers and colleagues from genetics and other departments: Gerben, Peter A, Christine, Hans, Yvonne, Ludolf, Dasha, Yurike, Tjakko, Yunia, Omid, Eva, Anna P., Mariska, Bahram, Jelkje, Jan O., Rudolf, Nicole, Klaas, Rolf, Anna D., Elena, Fany, Bine, Maria, Kaushal, Mateusz, Anna F., it was very nice knowing you. Dear Jana and Masha, thank you for a delicious sushi and nice conversations. Masha, I wish you a lot of success and I am sure, one day, you will become a great Manga Creator. Dear Mathieu B., thank you very much for taking care of my rabbit, Klapoucha. Good luck with your PhD study.

When I started my PhD in 2007, our Celiac group was rather small. Over these last five years, it was my pleasure to see how we grew. Dear Cisca, Aśka, Cleo, Isis, Gosia, Mathieu, Soesma, Barbara, Javier, Rodrigo, Vinod, Sebo, Senapati, Astrid, Patrick, Ania R., Maciek G., Sasha, and our old colleagues Monique, Ron, Martin W. and Marcel, thank you for all your input and comments during the Monday morning meetings. Dear super-Mathieu, I am still amazed by your energy. Thank you for always having time to go with me to the serum archive. Dear Soesma, I adore your children, especially your little girl. I will always remember your smile and high-heeled shoes. Droga Małgorzata T., Gosiu, dziękuję za mile spędzone chwile. Dear Isis, I was honored to be your supervisor. I know you will become a great scientist one day. Dear Sasha, I appreciated your help in explaining the basics of population genetics to me and I enjoyed our trip in California.

Dear colleges in the bioinformatics unit, thank you for your input for my work. Dear Gerard te Merman and Lude, thank you for trying to explain some basic statistics to me. Dear Freerk, Morris, Alex, Patrick, Marc Jan, Harm-Jan, Roan, Peter, Juha and Jorris, thank you for helping me with the next-generation sequencing data, the annotation and writing some helpful scripts for me. Dear Jingyuan, I would like to thank you for helping me with *R*. Dear Freerk and Robert W., I really enjoyed the drum and bass parties with you.

Dear IBD team - dear Rinse, you are for sure the coolest MD I have ever met. Dear Noortje, thank you for our good time together, the dinners and going out. It was nice to stay at your parents' home, since then I've been eating Clover honey for



breakfast. Karin, I think you are one of the most honest people I have ever known and because of that, it is so easy to communicate with you. Mitja, thank you for tons of movies, Balkan music and a fine time. Suzaaaaanneeeee, pumpkin, you are just one of the best roommates I have ever had.

My friends and buddies, it was a pleasure to share time with you, traveling, listening to music, drinking beer, relaxing, and discussing important as well as not so important issues :-).

Dimer, Albert, Remco, thank you for all the good fun, going out and time well spent in pubs :-). Marcinie, dziękuję Tobie za luncze, dowożone przez Aşkę.

My “must know” couple, Magda and Christian, I enjoyed staying at your place in Rotterdam and crazy outings in Warsaw.

Lotte and Ruud, thank you for being my buddies, especially at the beginning of my PhD. Dear Arnoud, thank you for helping me when you could.

Mats, my roommate, my friend, thank you for listening to all my troubles. I will always remember you and your passion for Indian food.

Monique, my friend, you are a very warm person. I enjoyed going with you to festivals and walking over the tidal mudflats (*wadlopen*). The Irish pub is where we should meet again. Dineke, isn't it funny that me (the drama queen) and you (a person who always eats bananas for lunch and old Amsterdam cheese) became rather good friends? Whenever I need some sense in my life and a bit of structure I will contact you. It was a pleasure seeing you growing as a PI and now I hope to see you growing as a mother. Thank you for your friendship.

Jihane, you are one of the sweetest girls I have ever met. Thank you for being next to me in the most difficult times as well as in the good ones. I hope to earn enough money to visit you one day in Lebanon. Kisses.

Céline, although you do not pronounce “h”, I still adore you and your “frunchness”. I really look forward to going to France with you.

Giulia, the first time we spoke we had an argument and I thought you would not like me ... but you proved me wrong. I don't know you that well, but I hope we will have a chance to become close friends as I really like you and appreciate your stubbornness :-).

Anno W., jak z nikim innym tylko z Tobą mogę rozmawiać o wszystkich sprawach sercowych i międzyludzkich. Wydaje mi się, że wiele nas łączy: charakter typu drama queen, wydawanie pieniędzy, podobne błędy życiowe i muzyka. Chyba dlatego dobrze się rozumiemy...i niech tak zostanie. Całuski.

Anno Rybak, Rybaku, co jak co, ale Opener tylko z Tobą. Jesteś bardzo pozytywną

osobą. Początki moje w Groningen były pełne śmiechu. Lubimy się i czerpiemy z tego wiele radości i to jest takie miłe :-).

Anno R. i Norbercie, cieszę się że udało nam się odnowić znajomość. Mile wspominałam waszą wizytę w Groningen.

Luizo kochana moja, zostawiłam Cię w Polsce ale nigdy nie zapomniałam o Tobie. Cieszę się, że odnalazłaś swoje szczęście w roli matki, żony i naukowca. Obiektywnie patrzysz na moje sprawy i zawsze jesteś ze mną szczerą, nigdy nie przestawaj tego robić. Dziękuję za miłość. Buziaki.

Kochana Miraim, Aśko, na początku było dużo muzyki i dobrej zabawy. Później wspólne wieczorki i obiadki (pierogi, nadziewane kaczki i zupa z pieca). A potem to już tylko bindole i Habibąki. Widzę wielki potencjał w lovebird, niech nam wzleci. Trzymam kciuki. Fajnie się z Tobą rozmawia, od rzeczy i do rzeczy, podróżuje i wydaje pieniądze na ciuchy, kosmetyki i inne takie ... Wiem że zawsze mogę na Ciebie liczyć i właśnie na tym polega przyjaźń. Dziękuję.

My dear Valentin, although we met just at the end of my stay in Groningen, you have already become one of the most important people in my life. You make me happy and I wish for myself to have many beautiful moments with you. I hope we can spend some more time together soon. Kisses :\*

Dziękuję swojej rodzinie. Drogi Krzyśku, dziękuję, że mogę na Ciebie liczyć w różnych sprawach życiowych: pielęgnacja ogrodu, pomoc w przeprowadzce, pomoc w zatrudnieniu i inne.

Kochana siostrzyczko, Zofio. Jestem z Ciebie bardzo dumna, że studiujesz, że masz pasję i że nieugięcie dążysz do swoich celów. Mam nadzieję, że się trochę nauczę od ciebie racjonalnego spojrzenia na świat i pewnego dystansu do życia.

Moja droga mamó. Zawsze byłaś przy mnie i zawsze mnie wspierałaś. Dziękuję za miłość i pomoc, tą materialną :-), ale przede wszystkim mentalną.

To all these fine people, I am grateful that I had the opportunity to meet you and I wish you all success and hope to see you again soon.

Thank you – Dank u wel – Dziękuję Wam!

Agata



Agata Maria Szperl was born on 16<sup>th</sup> June 1982, in Kielce, Poland. She followed the master's program in Biotechnology, specializing in Health Care and Animal Production, for four years at Warsaw Agricultural University, Poland. In the 4<sup>th</sup> year she went on an Erasmus exchange program to Wageningen University in the Netherlands. She was awarded a Wageningen University fellowship in 2005 to continue her MSc studies in Wageningen and performed the practical work for her MSc thesis in the Department of Microbial Physiology there. This research was aimed at isolating and characterizing anaerobic organisms involved in the anaerobic oxidation of methane. She graduated with an MSc in Medical Biotechnology from Wageningen University in 2007.

In November 2007 she was accepted for a PhD scholarship at the Department of Medical Genetics, University Medical Center Groningen, the Netherlands, to work in the research group of Prof. Cisca Wijmenga. During this period Agata was involved in a wide range of studies, including mapping causative genes in families segregating for complex and monogenic diseases, as well as in population-based studies on complex diseases like celiac disease and ulcerative colitis. During her PhD study, she went on a working visit to Dr. Paivi Saavalainen, Department of Medical Genetics, Biomedicum, Helsinki, Finland, for two months, financed by a Marie Curie grant. There she performed a replication study in ulcerative colitis cohort. Agata learned how to use and apply the newest sequencing technologies in order to find causative genes. In the future she would like to apply this knowledge to patient care and would therefore very much like to become a medical clinical geneticist.

Her research on the methods and applications for mapping causative genes in monogenic and complex diseases is described in this thesis.

