Genome-wide approaches towards identification of susceptibility genes in complex diseases

Genoomwijde strategieën tot de identificatie van ziekte veroorzakende genen in complexe ziekten

Lude Franke

346,054 letters from the DNA of Lude Franke

White	Homozygous wild-type allele
100% Metallic	Heterozygous
50% Metallic	Homozygous rare allele
Deletions	Deletions identified by TriTyper

Chapter 8 discusses the ethical issues of knowing these 346,054 letters.

30,000,000 bp

Deletion

20,000,000 bp

Genome-wide approaches towards identification of susceptibility genes in complex diseases

Genoomwijde strategieën tot de identificatie van ziekte veroorzakende genen in complexe ziekten

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector Magnificus, prof. dr. J.C. Stoof, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op dinsdag 27 mei 2008 des middags 2.30 uur

door

Lude Hendrikus Franke
geboren op 14 januari 1980 te Oldenzaal

Promotor

Prof. dr. C. Wijmenga

80,000,000 bp

Deletion

Preface

Many human diseases have a considerable genetic component and genome-wide research, initially through linkage analysis and, more recently, association analysis, has identified the genetic basis for various disorders, which has been useful for several reasons. First of all, these disease genes have been shown to be obvious starting points for functional follow-up research to learn more about the affected molecular biology underlying these diseases. As such, these affected pathways provide potential targets for pharmaceutical intervention. These findings have also been valuable in genetic counseling: by identifying the genetic basis of certain diseases, for example, cystic fibrosis[1], Duchenne's muscular dystrophy[2], and rare variants in BRCA1[3] and BRCA2[4] that confer a strong susceptibility to breast cancer, individuals who are suspected of being at risk have been enabled to make better-informed decisions for themselves and their offspring.

For various Mendelian diseases, genome-wide linkage analysis in families or sib-pairs and transmission-disequilibrium tests in trios (e.g. parents-child) have led to the identification of susceptibility loci (see figure 1A). Often, the subsequent positional cloning of the disease genes has led to the discovery of the precise genetic variants. Considerable efforts were later devoted to identifying susceptibility loci in more common, but also more complex, diseases. However, the results for these diseases were less conclusive and various explanations were given to explain the lack of success of genome-wide linkage studies. As the applied statistical models assumed certain models of inheritance and penetrance, it was questioned whether these assumptions applied to complex diseases as well. Another explanation was that complex diseases might be caused by numerous common variants, each of which conferred only a limited risk[5] (see figure 1B). However, the locations and characteristics of these common variants were still mostly unknown. Large collaborative projects (such as the Human Genome Project[6] and the International Haplotype Mapping Project[7; 8]) were instrumental in providing this knowledge. Technological improvements have enabled the use of genome-wide DNA oligonucleotide arrays that can now capture most of the germline heritable common SNP variation.

The principal aim of this thesis was to identify new genes in complex diseases using these genome-wide approaches. **Part 1** describes new statistical methods developed to help identify susceptibility variants. We also applied genome-wide strategies to explain the functional consequences of some of these variants. **Part 2** describes two genome-wide association studies we performed that led to identification of susceptibility loci in celiac disease and amyotrophic lateral sclerosis. **Part 3** discusses our results, outlines some future perspectives, and explores the consequences these findings are having in the field of commercial genetic testing services.

1 Rommens JM, Iannuzzi MC, Kerem B, Drumm ML, Melmer G, Dean M, Rozmahel R, Cole JL, Kennedy D, Hidaka N, *et al.* (1989) Identification of the cystic fibrosis gene: chromosome walking and jumping. Science (New York, NY 245:1059-1065)

2 Hoffman EP, Brown RH, Jr., Kunkel LM (1987) Dystrophin: the protein product of the Duchenne muscular dystrophy locus. Cell 51:919-928

3 Futreal PA, Liu Q, Shattuck-Eidens D, Cochran C, Harshman K, Tavtigian S, Bennett LM, Haugen-Strano A, Swensen J, Miki Y, *et al.* (1994) BRCA1 mutations in primary breast and ovarian carcinomas. Science (New York, NY 266:120-122)

4 Wooster R, Neuhausen SL, Mangion J, Quirk Y, Ford D, Collins N, Nguyen K, Seal S, Tran T, Averill D, *et al.* (1994) Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. Science (New York, NY 265:2088-2090)

5 Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science (New York, NY 273:1516-1517)

6 Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, *et al.* (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921

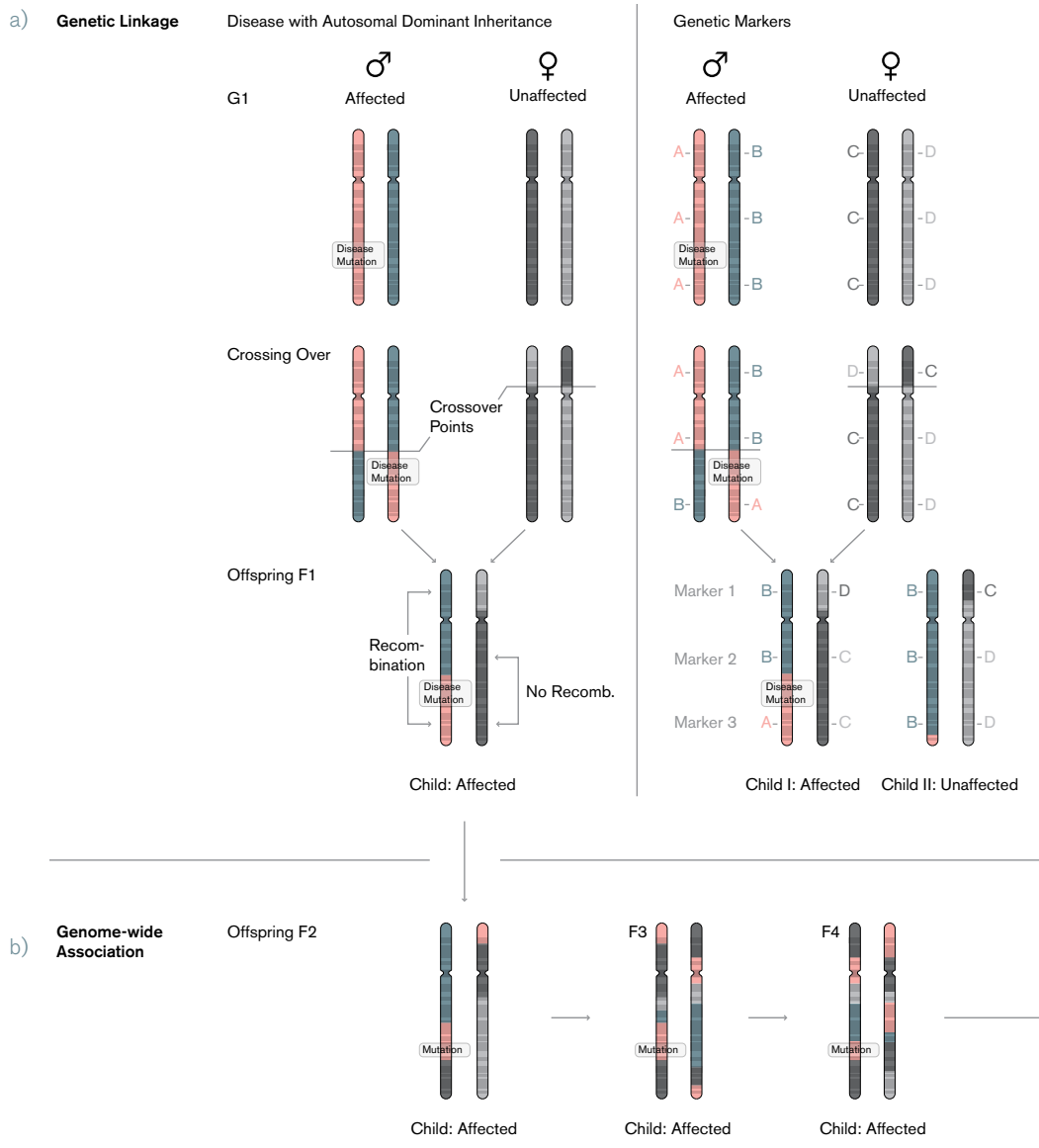
7 International HapMap Consortium, (2005) A haplotype map of the human genome. Nature 437:1299-1320

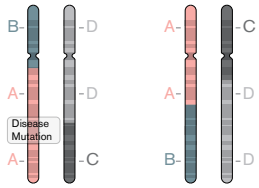
8 Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851-861

Figure 1

Concept of linkage analysis and genome-wide association analysis.

a) Linkage analysis involves the analysis of multiple related individuals. In this example a male (generation G1) is affected by a mutation on one haplotype (indicated by red) that has an autosomal dominant inheritance pattern. Children I and III have inherited the disease locus. Subsequent typing of genetic markers, spaced equally over this chromosome, allows for determining which markers are linked with the disease (Marker 2 in this example). Indicated on the right are the characteristics of both linkage and genome-wide association analysis. **b)** Genome-wide association analysis uses unrelated individuals, but assumes many generations ago individuals had common ancestors. In this example we assume a founder mutation in generation G1. It is evident that in subsequent generations the disease causing haplotype becomes smaller. If one would compare cases and controls from the current generation (generation F6), many markers are necessary to identify the disease associated haplotype.





Child III: Affected Child IV: Unaffected

Genetic Linkage Analysis

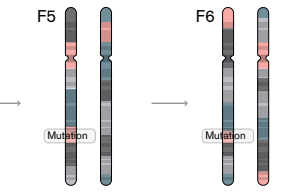
Assumptions In This Example:
- Random Mating, thus Hardy-Weinberg Equilibrium
- Mendelian segregation
- Female and Male Recombination Rates Indentical
- Full Penetrance, No Phenocopies

Strengths of Genetic Linkage Analysis:
- Not sensitive to allelic heterogeneity or population history
- Applicable to limited number of samples: inexpensive

Weaknesses of Genetic Linkage Analysis:
- Sensitive to genetic heterogeneity (e.g. multiple loci)
- Allele frequency assumed
- Affection status correctly assumed (Few phenocopies)
- Mapped locus size usually large

Alleles Inherited from Affected G1	Number of Alleles Observed In Affected Children A	Affected Children B	Unaffected Children A	Unaffected Children B
⟶ Marker 1	1	1	1	1
⟶ Marker 2	1	1	1	1
⟶ Marker 3	2	0	0	2

Conclusion: Disease Maps in Vicinity of Marker 3



Child: Affected Child: Affected

Genome-wide Association Analysis

Assumptions in This Example:
- Founder mutation in generation G1
- Cases and controls sampled from generation F6
- Case and control allele frequencies compared

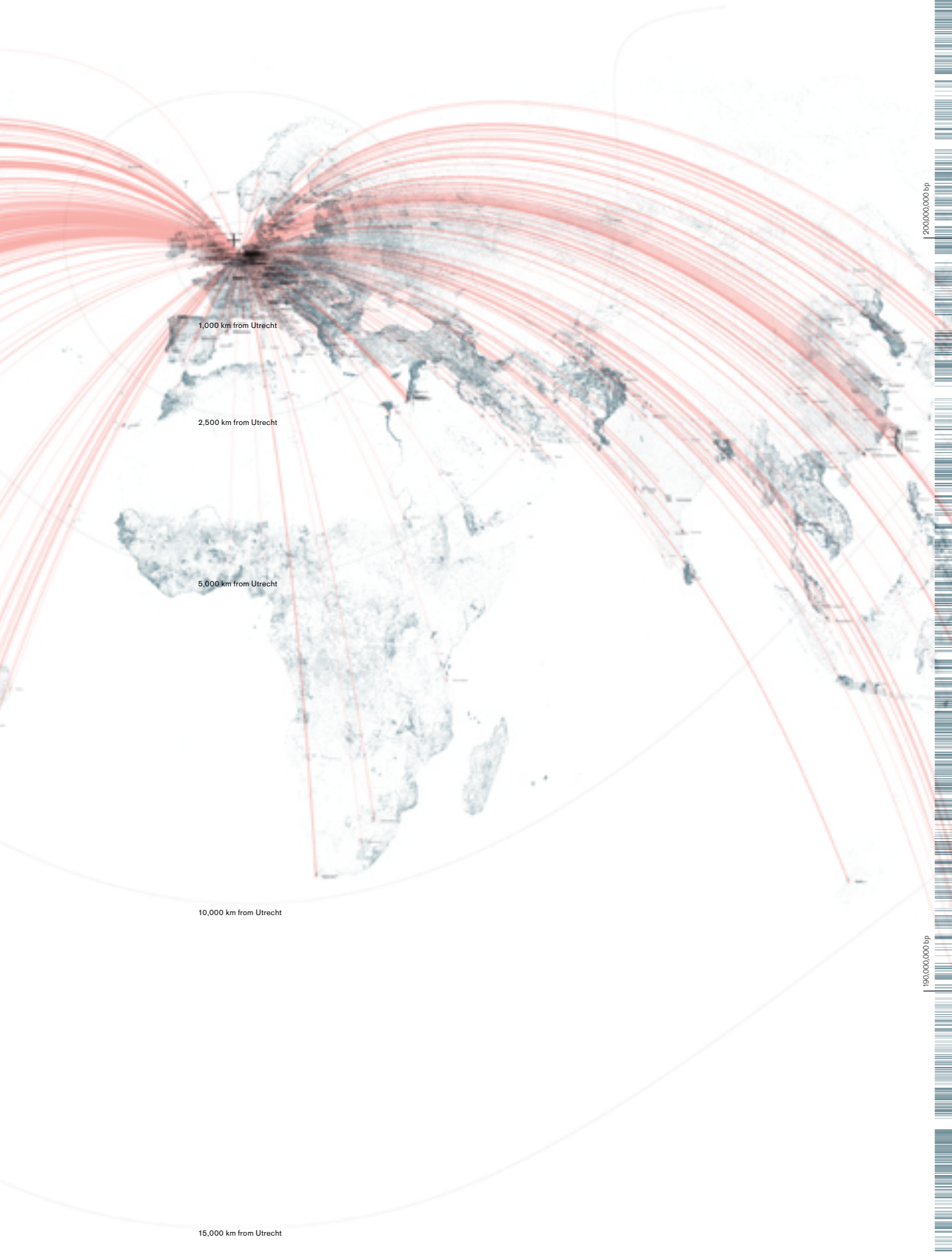
Consequences of more subsequent generations:
- More markers are necessary to map locus
- Mapped locus size smaller
- Still linkage disequilibrium present between nearby loci

Visits to www.genenetwork.nl and www.prioritizer.nl
Since their inception in 2006, www.genenetwork.nl and www.prioritizer.nl have been visited 17,147 times.

Most of the visitors have been from Europe and the USA. If all the visitors want to meet, the gathering should take place in the middle of the North Sea (indicated with a cross). This coordinate represents the least traveling that would be necessary.

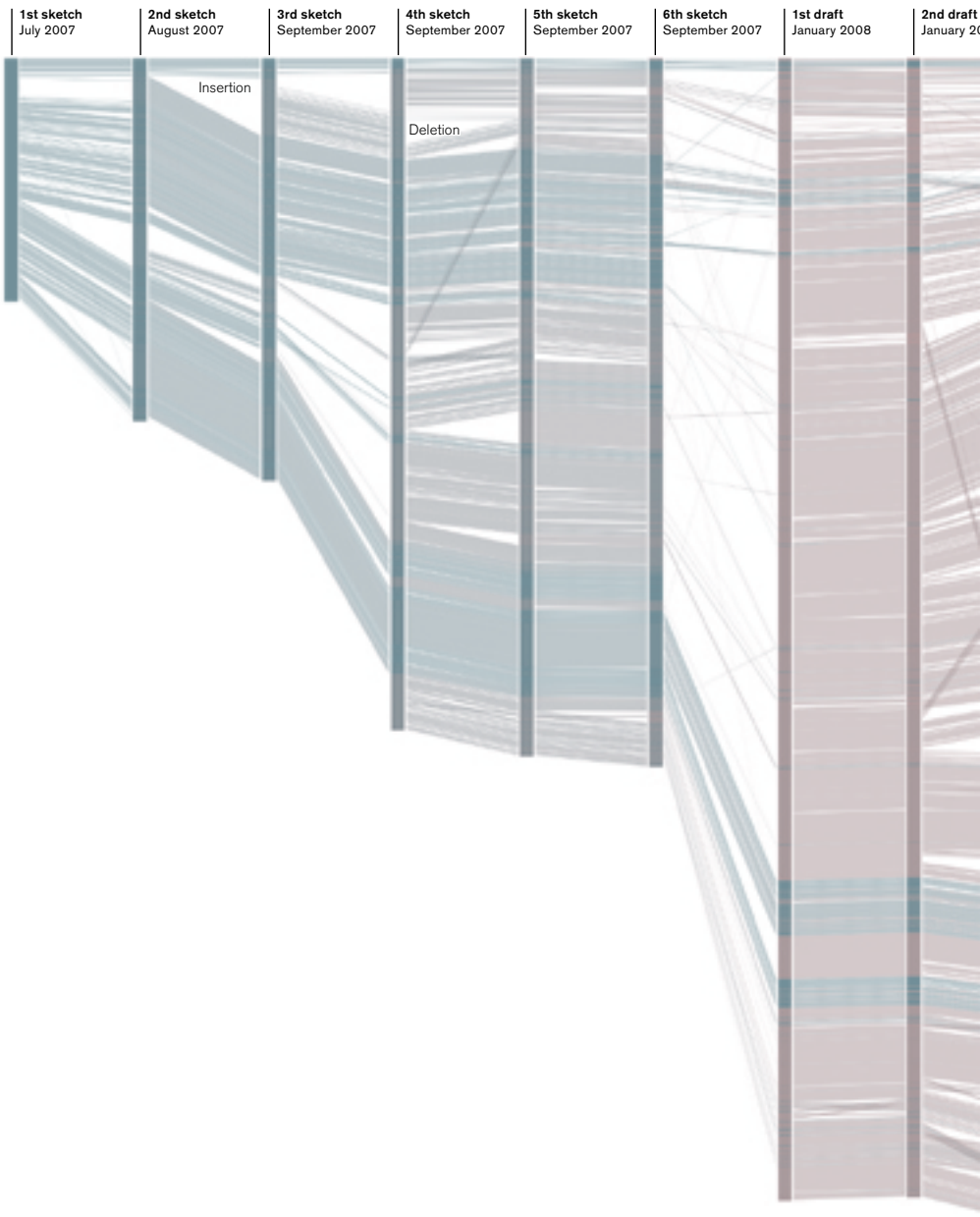
2,673,765 known cities and towns (each with over 1,000 inhabitants) are represented by individual pixels (source: geobytes.com). Contour lines indicate the distance from Utrecht. Projection used: Winkel-Tripel.

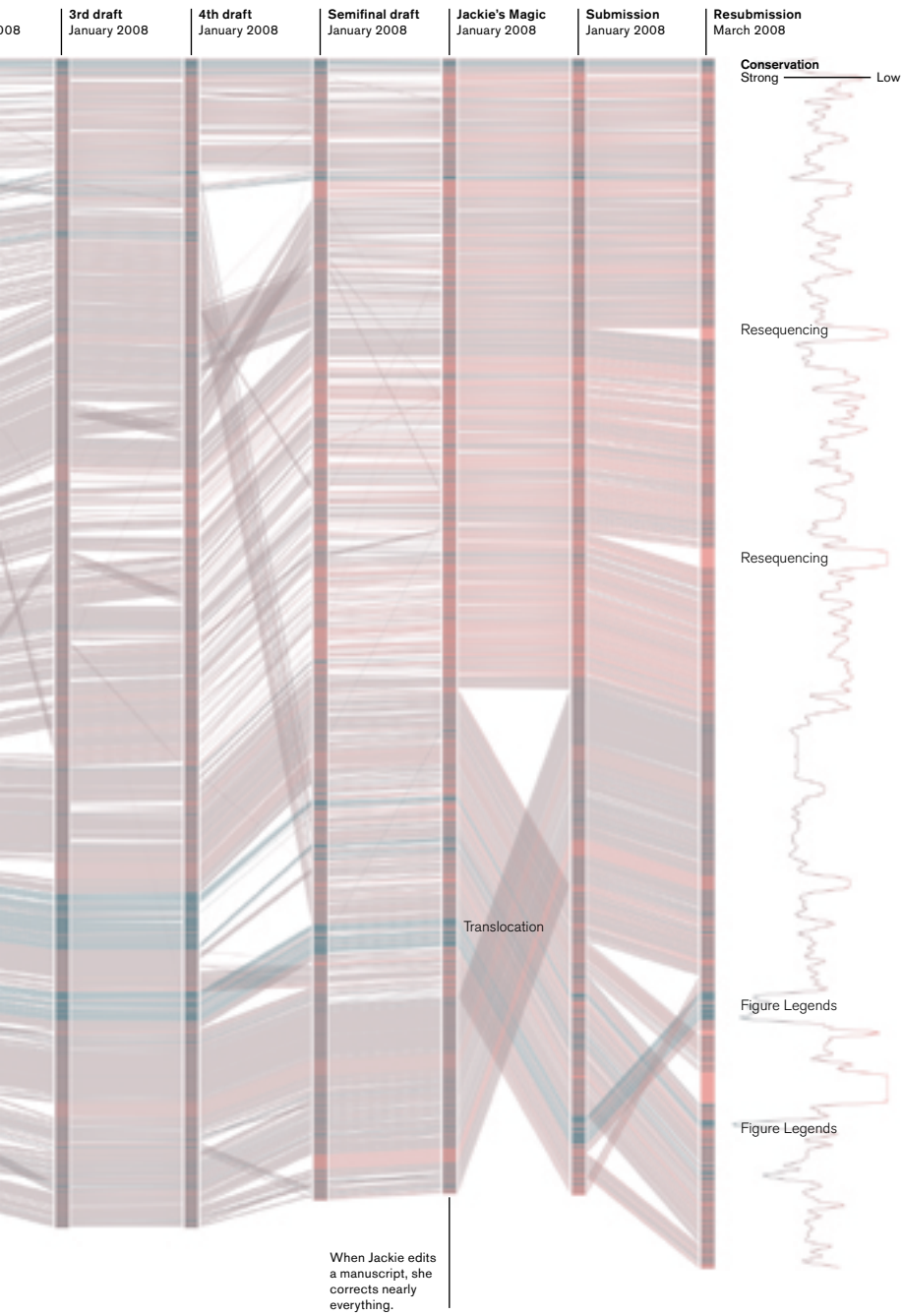
Scientists in Cuba have visited our sites, but no one from North Korea has made a visit.


1,000 km from Utrecht
2,500 km from Utrecht
5,000 km from Utrecht
10,000 km from Utrecht
15,000 km from Utrecht
200,000,000 bp
190,000,000 bp

Evolution of chapter 3

Writing a manuscript takes time and during its course and along its pathway from concept to publication, a considerable amount of text is introduced, edited, rearranged or deleted.




008
3rd draft
January 2008
4th draft
January 2008
Semifinal draft
January 2008
Jackie's Magic
January 2008
Submission
January 2008
Resubmission
March 2008
Conservation
Strong
Low
Resequencing
Resequencing
Translocation
Figure Legends
Figure Legends
When Jackie edits a manuscript, she corrects nearly everything.


230,000,000 bp
Deletion
220,000,000 bp

E-mail received from colleagues during my PhD period
This infograph shows the history of 3,564 e-mails received from colleagues.

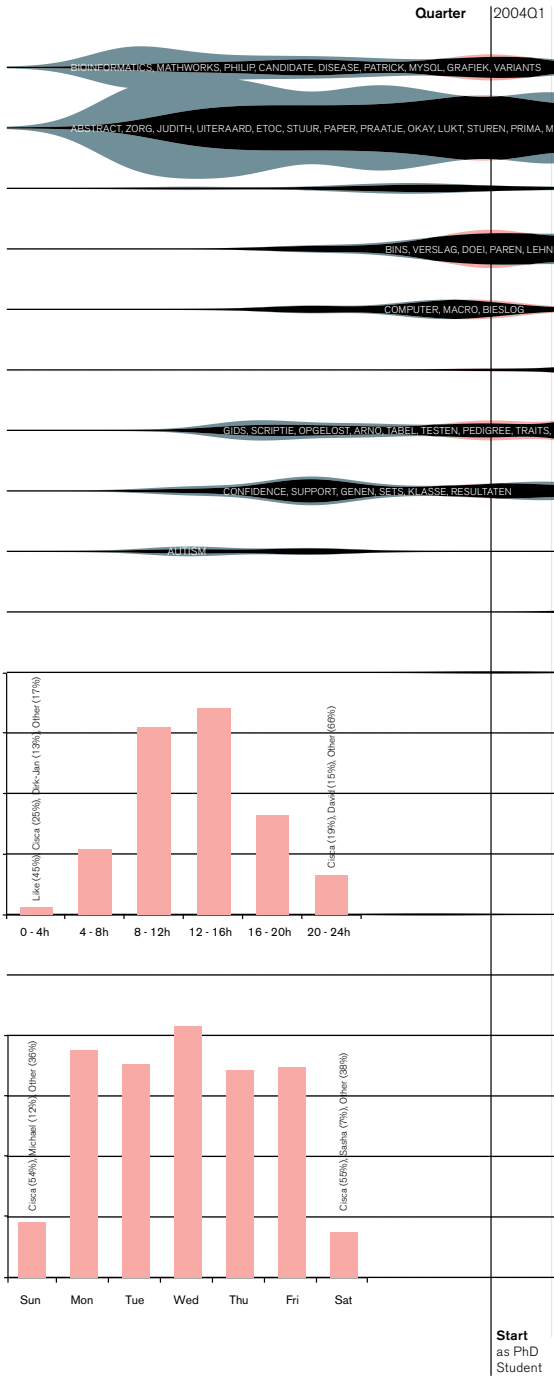
Red indicates the absolute number of e-mails received, blue indicates the relative amount of e-mail received per individual compared to all my other colleagues. White texts per individual denote a selection of significantly overrepresented words (Fisher's exact test, $P < 10^{-6}$).

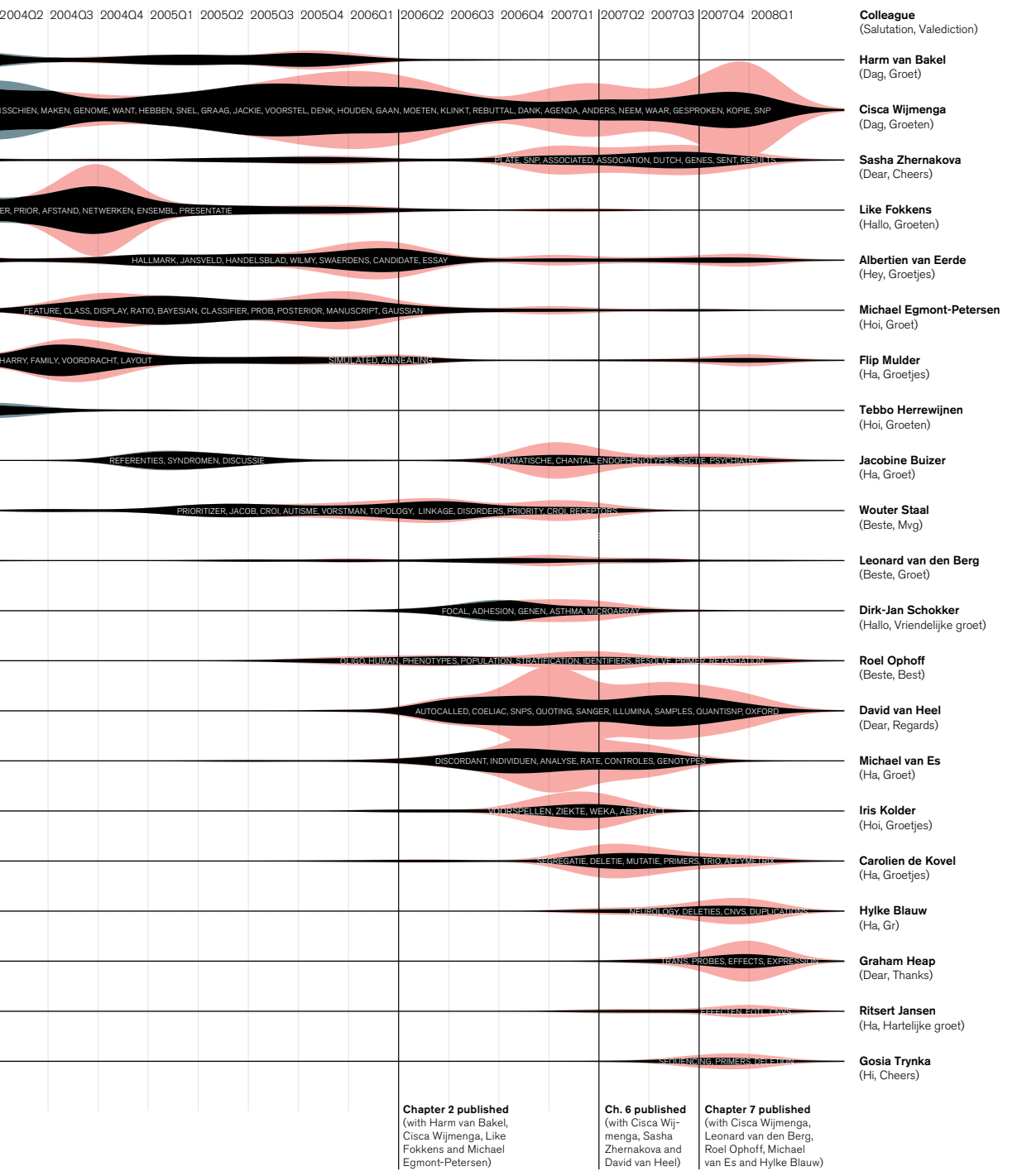
Publication times of major articles are indicated, along with the people who contributed to these papers.

Cisca Wijmenga is the overall winner, both in the relative and absolute numbers of e-mails received from her. She regularly sent e-mail during the night or weekends.

Like Fokkens, Michael Egmont-Petersen, Dirk-Jan Schokker and Sasha Zhernakova also liked to send e-mail in the evening and weekends.

Most people seem to use a fixed salutation and valediction, but these differ considerably from person to person.




2004Q2
2004Q3
2004Q4
2005Q1
2005Q2
2005Q3
2005Q4
2006Q1
2006Q2
2006Q3
2006Q4
2007Q1
2007Q2
2007Q3
2007Q4
2008Q1
Colleague
(Salutation, Valediction)
Harm van Bakel
(Dag, Groet)
SSCHIEN, MAKEN, GENOME, WANT, HEBBEN, SNEL, GRAAG, JACKIE, VOORSTEL, DENK, HOUDEN, GAAN, MOETEN, KLINKT, REBUTTAL, DANK, AGENDA, ANDERS, NEEM, WAAR, GESPROKEN, KOPIE, SNP
Cisca Wijmenga
(Dag, Groeten)
PLATE, SNP, ASSOCIATED, ASSOCIATION, DUTCH, GENES, SENT, RESULTS
Sasha Zhernakova
(Dear, Cheers)
ER, PRIOR, AFSTAND, NETWERKEN, ENSEMBL, PRESENTATIE
Like Fokkens
(Hallo, Groeten)
HALLMARK, JANSVELD, HANDELSBLAD, WILMY, SWAERDENS, CANDIDATE, ESSAY
Albertien van Eerde
(Hey, Groetjes)
FEATURE, CLASS, DISPLAY, RATIO, BAYESIAN, CLASSIFIER, PROB, POSTERIOR, MANUSCRIPT, GAUSSIAN
Michael Egmont-Petersen
(Hoi, Groet)
HARRY, FAMILY, VOORDRACHT, LAYOUT
SIMULATED, ANNEALING
Flip Mulder
(Ha, Groetjes)
Tebbo Herrewijnen
(Hoi, Groeten)
REFERENTIES, SYNDROMEN, DISCUSSIE
AUTOMATISCHE, CHANTAL, ENDOPHENOTYPES, SECTIE, PSYCHIATRY
Jacobine Buizer
(Ha, Groet)
PRIORITIZER, JACOB, CROI, AUTISME, VORSTMAN, TOPOLOGY, LINKAGE, DISORDERS, PRIORITY, CROI, RECEPTORS
Wouter Staal
(Beste, Mvg)
Leonard van den Berg
(Beste, Groet)
FOCAL, ADHESION, GENEN, ASTHMA, MICROARRAY
Dirk-Jan Schokker
(Hallo, Vriendelijke groet)
OLIGO, HUMAN, PHENOTYPES, POPULATION, STRATIFICATION, IDENTIFIERS, RESOLVE, PRIMER, RETARDATION
Roel Ophoff
(Beste, Best)
AUTOCALLED, COELIAC, SNPS, QUOTING, SANGER, ILLUMINA, SAMPLES, QUANTISNP, OXFORD
David van Heel
(Dear, Regards)
DISCORDANT, INDIVIDUEN, ANALYSE, RATE, CONTROLES, GENOTYPES
Michael van Es
(Ha, Groet)
VOORSPELLEN, ZIEKTE, WEKA, ABSTRACT
Iris Kolder
(Hoi, Groetjes)
SEGREGATIE, DELETIE, MUTATIE, PRIMERS, TRIO, AFFYMETRIX
Carolien de Kovel
(Ha, Groetjes)
NETBIOLOGY, DELETIES, CNVS, DUPLICATIONS
Hylke Blauw
(Ha, Gr)
TRANS, PROBES, EFFECTS, EXPRESSION
Graham Heap
(Dear, Thanks)
Ritsert Jansen
(Ha, Hartelijke groet)
SEQUENCING, PRIMERS, DELETION
Gosia Trynka
(Hi, Cheers)
10,000,000 bp
Chapter 2 published
(with Harm van Bakel, Cisca Wijmenga, Like Fokkens and Michael Egmont-Petersen)
Ch. 6 published
(with Cisca Wijmenga, Sasha Zhernakova and David van Heel)
Chapter 7 published
(with Cisca Wijmenga, Leonard van den Berg, Roel Ophoff, Michael van Es and Hylke Blauw)

Collaborations

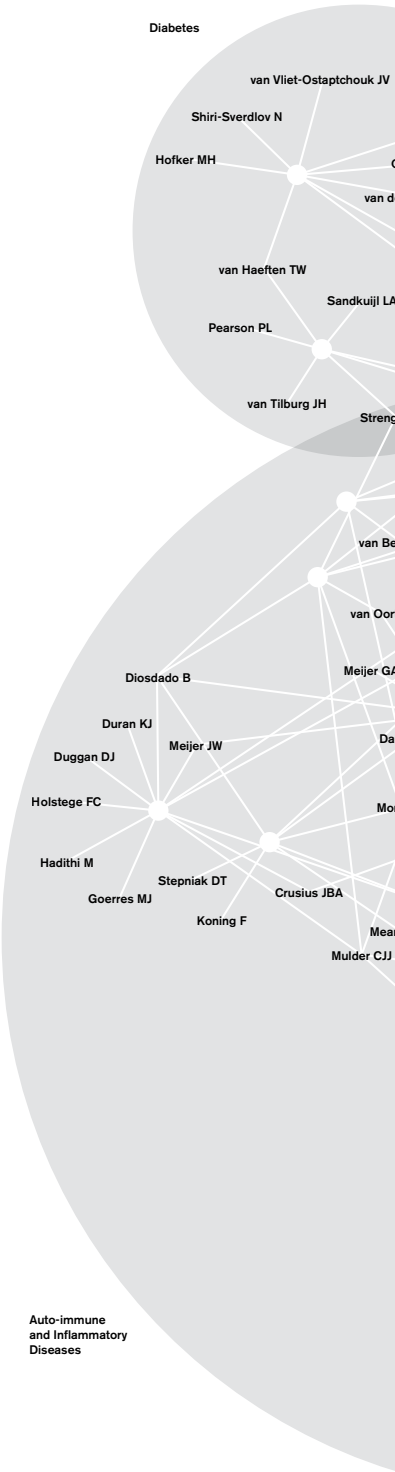
As is evident from this infographic, science has become highly collaborative. Shown are all co-authors on publications.

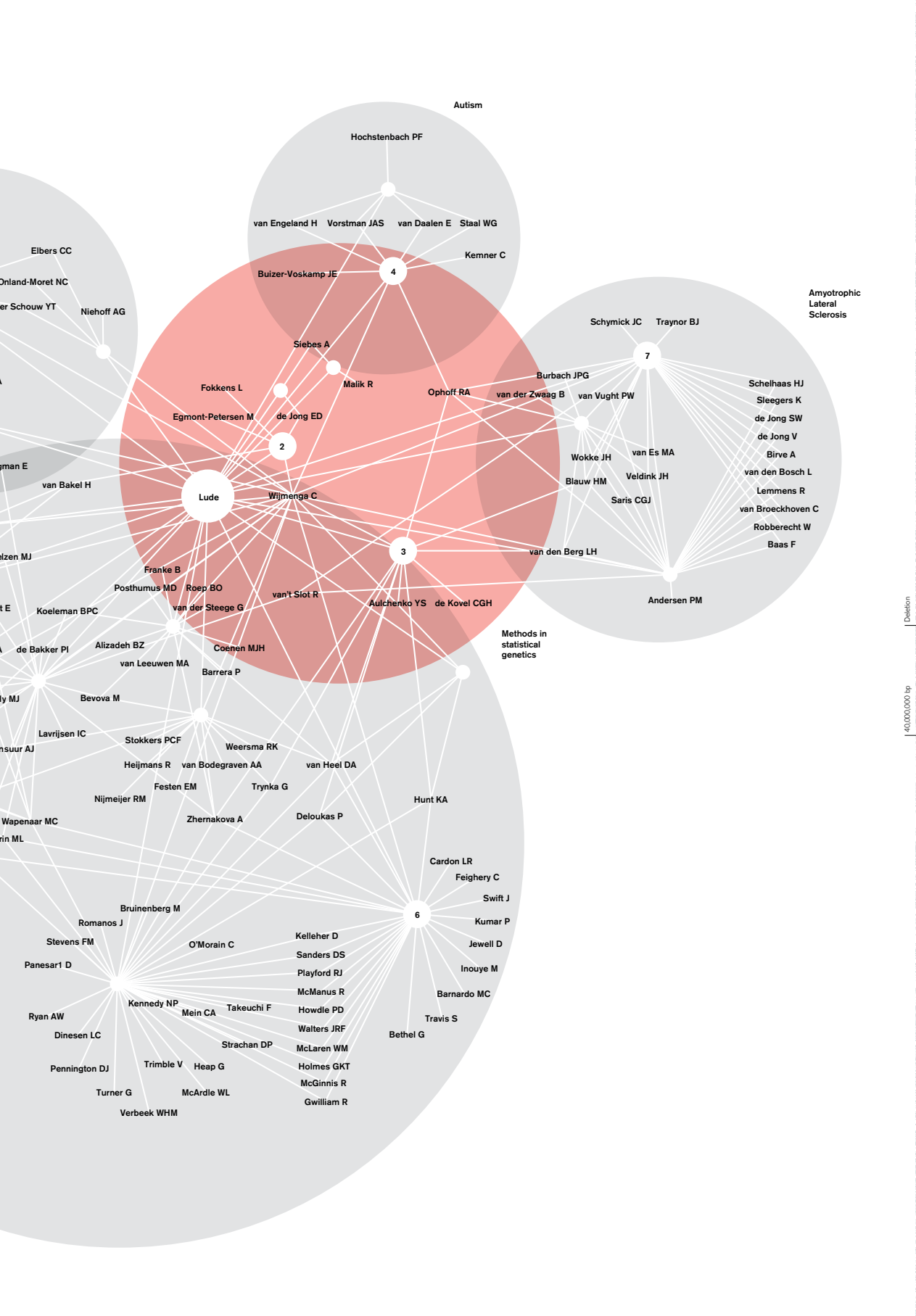
Small white circles denote publications. If these publications are included in this thesis, a number within the circle refers to the corresponding chapter.

Most of the articles reflect disease specific studies. Indicated in red are five publications that are methodological.

Although with exceptions, a correlation is present between the amount of co-authors and the impact factor of the journals in which these papers have been published.

Cisca Wijmenga, Martin Wapenaar, Sasha Zhernakova and David van Heel have co-authored most of the papers.




Autism
Hochstenbach PF
van Engeland H
Vorstman JAS
van Daalen E
Staal WG
Kemner C
Buizer-Voskamp JE
4
Elbers CC
Onland-Moret NC
er Schouw YT
Niehoff AG
Amyotrophic Lateral Sclerosis
Schymick JC
Traynor BJ
7
Siebes A
Malik R
Fokkens L
Ophoff RA
Burbach JPG
van der Zwaag B
van Vught PW
Schelhaas HJ
Sleegers K
de Jong SW
de Jong V
Birve A
van den Bosch L
Lemmens R
van Broeckhoven C
Robberecht W
Baas F
Egmont-Petersen M
de Jong ED
2
Wokke JH
van Es MA
Veldink JH
Blauw HM
Saris CGJ
van Bakel H
Lude
Wijmenga C
3
van den Berg LH
Franke B
Posthumus MD
Roep BO
van't Slot R
Aulchenko YS
de Kovel CGH
Andersen PM
Koeleman BPC
van der Steege G
Alizadeh BZ
de Bakker PI
Coenen MJH
van Leeuwen MA
Barrera P
Methods in statistical genetics
Bevova M
Lavrijsen IC
Stokkers PCF
Weersma RK
Heijmans R
van Bodegraven AA
van Heel DA
Festen EM
Trynka G
Nijmeijer RM
Hunt KA
Zhernakova A
Deloukas P
Cardon LR
Feighery C
Swift J
Bruinenberg M
6
Kumar P
Romanos J
Stevens FM
O'Morain C
Kelleher D
Jewell D
Sanders DS
Panesar1 D
Playford RJ
Inouye M
McManus R
Barnardo MC
Kennedy NP
Takeuchi F
Howdle PD
Ryan AW
Mein CA
Travis S
Dinesen LC
Walters JRF
Bethel G
Strachan DP
McLaren WM
Trimble V
Heap G
Holmes GKT
Pennington DJ
McGinnis R
Turner G
McArdle WL
Gwilliam R
Verbeek WHM
Deletion
40,000,000 bp

2 Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes

American Journal of Human Genetics, 2006 Jun;78(6): 1011-25

Lude Franke[1], Harm van Bakel[1], Like Fokkens[1], Edwin D. de Jong[2], Michael Egmont-Petersen[3], Cisca Wijmenga[1]

1 Complex Genetics Section, Department of Biomedical Genetics–Department of Medical Genetics, University Medical Centre Utrecht
2 Large Distributed Databases Group, Institute of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands
3 Department of Human Genetics, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

Summary

Most common genetic disorders have a complex inheritance and may result from variants in many genes, each contributing only weak effects to the disease. Pinpointing these disease genes within the myriad of susceptibility loci identified in linkage studies is difficult because these loci may contain hundreds of genes. However, in any disorder, most of the disease genes will be involved in only a few different molecular pathways. If we know something about the relationships between the genes, we can assess whether some genes (which may reside in different loci) functionally interact with each other, indicating a joint basis for the disease etiology. There are various repositories of information on pathway relationships. To consolidate this information, we developed a functional human gene network that integrates information on genes and the functional relationships between genes, based on data from the Kyoto Encyclopedia of Genes and Genomes, the Biomolecular Interaction Network Database, Reactome, the Human Protein Reference Database, the Gene Ontology database, predicted protein-protein interactions, human yeast two-hybrid interactions, and microarray coexpressions. We applied this network to interrelate positional candidate genes from different disease loci and then tested 96 heritable disorders for which the Online Mendelian Inheritance in Man database reported at least three disease genes. Artificial susceptibility loci, each containing 100 genes, were constructed around each disease gene, and we used the network to rank these genes on the basis of their functional interactions. By following up the top five genes per artificial locus, we were able to detect at least one known disease gene in 54% of the loci studied, representing a 2.8-fold increase over random selection. This suggests that our method can significantly reduce the cost and effort of pinpointing true disease genes in analyses of disorders for which numerous loci have been reported but for which most of the genes are unknown.

Introduction

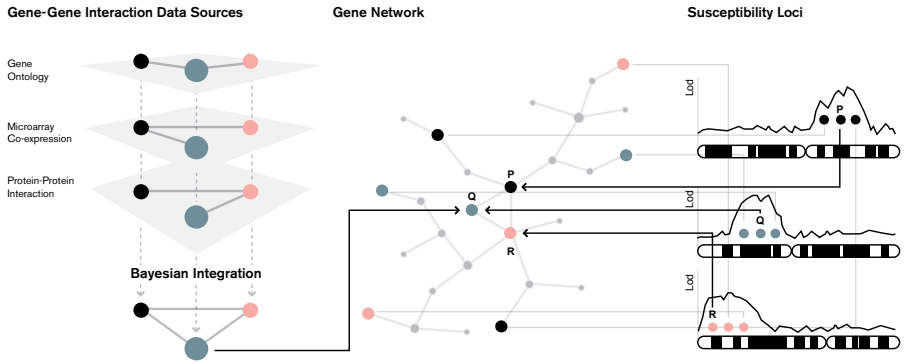
The completion of various genome-sequencing projects and large-scale genomic studies has led to a wealth of available biological data. It is anticipated that this information will revolutionize our insight into the molecular basis of most common diseases by making it easier and quicker to identify genes with variants that predispose to disease (i.e., disease genes). At the moment, we are faced with many disease susceptibility loci, resulting from linkage or cytogenetic analyses, that cover extensive genomic regions. Usually, when the genes in these loci are assessed, positional candidate genes become apparent that can be linked to the phenotype being studied on the basis of their biological function. However, the most obvious functional candidate gene from a disease locus does not always prove to be involved in the disease[e.g.1–5]. Often, genes that would not have been predicted to be disease causing prove to be the true disease gene—for example, the BRCA1 gene in early-onset breast cancer[6]. Moreover, although these disease genes might have been assigned biological functions, it is not always evident how these functions relate to disease. Finally, genes with unknown functions are often overlooked, as attention is paid only to well-studied genes for which functions and interactions have been identified or implicated, some of which can be related to the disease pathogenesis. For example, in Fanconi anemia, at least 10 disease genes were identified[7], but only a few had a known function. However, follow-up research[8–10] revealed that five of those genes function in the same protein complex. Another example is limbgirdle muscular dystrophy, in which many of the disease genes encode for proteins that are part of the dystrophin complex[11]. This emphasizes the importance of taking an unbiased approach to assessing positional candidate genes.

Faced with the absence of complete functional information for the majority of genes in susceptibility loci, it is difficult to prioritize the positional candidate genes correctly for further sequence or association analysis. However, high-throughput genomic work has now yielded relatively unbiased genomewide data sets[12–15] that comprise known metabolic, regulatory, functional, and physical interactions. There is, however, little integration of these diverse data sets into a coherent view of possible gene and protein interactions that can be used to investigate relationships between genes in different genetic loci. We have tried to address this problem by developing a functional human gene network that comprises known interactions derived from the Biomolecular Interaction Network Database (BIND)[12], the Human Protein Reference Database (HPRD)[13], Reactome[15], and the Kyoto Encyclopedia of Genes and Genomes (KEGG)[14].

Since these data sets contain a limited number of known interactions, we implemented a Bayesian framework to complement these relationships with a large number of predicted interactions by relying on evidence for putative gene relationships based on biological process and molecular function annotations from the Gene Ontology database (GO)[16]. We further incorporated experimental data—namely, coexpression data derived from ~450 microarray hybridizations from the Stanford Microarray Database (SMD)[17] and the NCBI Gene Expression Omnibus (GEO)[18], along with human yeast two-hybrid (Y2H) interactions[19] and interactions based on orthologous high-throughput protein-protein interactions from lower eukaryotes[20]. Our interaction network was then used to test whether we could rank the best positional candidates in susceptibility loci on the basis of their interactions, assuming that the causative genes for any one disorder will be involved in only a few different biological pathways. This would be apparent in our network as a clustering of genes from different susceptibility loci, resulting in shorter gene-gene connections between disease genes than one would expect by chance (fig. 1). Our method (called "*Prioritizer*") analyzes susceptibility loci and investigates whether genes from different loci can be linked to each other directly[21] or indirectly[22]. When we constructed artificial loci of varying size around susceptibility loci from 96 different genetic disorders (each containing at least three loci) and used Prioritizer in our most comprehensive gene network to rank the positional candidate genes for each locus, we were able to significantly increase the chance of detecting disease genes.

Figure 1

Basic principles of the prioritization method for positional candidate genes with the use of a functional human gene network. The method integrates different gene-gene interaction data sources in a Bayesian way (left panel). Subsequently, this gene network is used to prioritize positional candidate genes, with all genes assigned an initial score of zero. In the example (right panel), three different susceptibility loci are analyzed, each containing a disease gene (P, Q, or R) and two nondisease genes. In each locus, the three positional candidate genes increase the scores of nearby genes in the gene network, by use of a kernel function that models the relationship between gene-gene distance and score effect. Genes within each locus are ranked on the basis of their eventual effect score, corrected for differences in the topology of the network (see the "Material and Methods" section).



Material and Methods

Functional Gene Network Reconstruction

As a basis for the gene network, we used annotations from Ensembl[23], version 32.35, resulting in 20,334 known genes that physically map within the autosomes or chromosome X or Y. This yielded 206,725,611 potential gene-gene interactions.

On the basis of this set of genes, a comprehensive "gold standard" set of validated direct gene-gene relationships (true positives) was determined using both BIND (September 15, 2005) and HRPD (September 15, 2005) to extract human, curated protein-protein interactions, the proteins of which were mapped to Ensembl gene identifiers. In addition, all pathways from Reactome (September 15, 2005) and KEGG (September 15, 2005) were used to derive direct interactions that were of transcriptional, physical, or metabolic origin, since pathways are usually composed of genes and proteins that interact with each other in various ways. We chose to allow interactions of physical, metabolic, and regulatory origin to be included within our network, because, for instance, mutations in either one of two genes encoding proteins in the same metabolic pathway or protein complex could lead to the same disease phenotype.

Because the true-positive gold standard only describes a limited number of relationships between a limited number of genes, we also used data from GO, coexpression data derived from microarray experiments, conserved protein-protein highthroughput data, and human Y2H interaction data to predict interactions of the remaining gene pairs. We used a Bayesian classifier, because these four types of data were of varying reliability and only contained information about a subset of the data. The classifier allows for combining dissimilar data sets, can deal with missing data, and uses conditional probabilities that can be well interpreted and that control for the varying reliability of the data sets[24–29].

Training of Bayesian Classifier on Gold Standard

For the prediction of interactions, we used a Bayesian classifier type that assumed all data sets had been binned. This operation was performed for each gene pair, and it determined, for each data set, to which bin the pair belongs. Because the number of bins per data set was limited, each bin contained many gene pairs. Subsequently, for each bin, we determined the likelihood ratio between the proportion of gene pairs known to interact and the proportion of gene pairs known not to interact. This measure indicates whether there is an overor an underrepresentation of truly interacting gene pairs in the bin, which specifies the conditional probability estimates of the Bayesian classifier; thus, training of the classifier is straightforward.

However, to be able to train the classifier by determining likelihood ratios of sets of gene pairs, it was crucial that the gold standard, containing the aforementioned well-defined set of curated true-positive gene pairs, be complemented with a set of gene pairs for which there is strong evidence that they, or the proteins they encode, do not functionally interact (true negatives). As has been discussed by others[30], the construction of this true-negative reference set is problematic, because it is impossible to be certain that two genes (i.e., their protein products) do not interact. However, by assuming that genes encoding for proteins localized within different cellular compartments are, in general, unrelated, it is possible to make a list of gene pairs that are unlikely to interact. The GO Cellular Component annotations were used to yield groups of gene pairs that have exclusive cellular component annotations. To overcome a strong selection bias in the classifier toward wellannotated genes (details provided in appendix A), only the 5,105 genes that were part of a true-positive gene pair at least three times were allowed to form true-negative gene pairs. We chose combinations of cellular organelles that were highly underrepresented ($\chi^2 = 2{,}490$; $P < 10^{-10}$) within the true-positive set, which resulted in gene pairs for the following combinations: nucleus and extracellular matrix, protein complex and Golgi apparatus, protein complex and Golgi stack, non–mem-

brane-bound organelle and Golgi stack, non–membrane-bound organelle and extracellular space, non– membrane-bound organelle and Golgi apparatus, extracellular region and organelle membrane, mitochondrion and extracellular matrix, extracellular space and organelle membrane, extracellular space and Golgi stack, organelle membrane and extracellular matrix, extracellular matrix and Golgi stack, extracellular matrix and ubiquitin ligase complex, and ubiquitin ligase complex and Golgi stack.

Preprocessing and Binning of Data Sets

To allow for Bayesian integration, the GO data, microarray coexpression data, and orthologous and human protein-protein interactions data were preprocessed and binned. Biological Process and Molecular Function GO annotations were derived from Ensembl, and two measures of relatedness for each of the two data sets were determined, resulting in a total of four different GO measures of relatedness. First, we determined, for each Biological Process GO term, how many of the genes had been assigned this term. Then, we determined which Biological Process GO terms were shared between the two components of each gene pair, for all the pairs. This led to the shared GO term that was annotated in the least number of genes, and its frequency of occurrence was used as a measure. GO terms GO:0000004 (biological process unknown) and GO:0005554 (molecular function unknown) were discarded, since genes that shared either of these highly unspecific terms should not be related to each other on the basis of this information. The same procedure was performed to generate the first measure of Molecular Function GO relatedness. The second measure determined the maximal hierarchical depth at which a gene pair shared a Biological Process GO term. This hierarchical depth was defined as the shortest number of branches necessary to go from one Biological Process GO term back to the GO root. The same method was used to generate the maximum hierarchical depth of the Molecular Function GO sharing measure.

Coexpression between genes was determined in microarray data sets from GEO and SMD. Individual data sets comprised an experiment that contained at least 10 hybridizations. To ensure that the quality of the intensity measurements was reliable, various filtering steps were performed to exclude spots with low signal-to-noise ratios[31]. Within the SMD data sets, intensity spots were filtered out that were either missing or contaminated, and the mean intensity of spots had to be at least 2.5 times higher than the average background signal of the microarray. Since GEO contains both ratiometric and Affymetrix single-spot intensity microarray data sets, we used different filtering strategies. The 5% of genes with the lowest maximal intensity were removed from the Affymetrix data sets. For both SMD and GEO, expression ratios were $\log_2$ transformed. Microarray features missing at least 25% of expression measurements in a data set after filtering were excluded. All features were assigned Ensembl gene identifiers by comparing their sequences to Ensembl transcripts with the use of SSAHA[32].

To determine which gene pairs showed coexpression, the mutual information was calculated between all the genes represented within each data set[33] if there were at least 10 nonmissing data points. As a preprocessing step, expression levels were ranked; this invertible reparameterization did not affect the mutual information. Next, for each pair of genes, the joint distribution of expression levels was estimated by calculating a histogram with overlapping windows. The range was divided into six windows, where each window extends to the center of the next window. The number of windows was chosen by optimizing the error rate for the mutual information derived from analytical probability densities[33]. In this way, each data point contributes to two windows, except at the extremities. Finally, on the basis of the resulting distribution, the mutual information (MI) between each pair of genes was calculated as $MI(A,B) = H(A) + H(B) - H(A,B)$, where $H(X)$ is the information- theoretic Shannon entropy[34]. For each microarray data set, the MI score was binned. This allowed the subsequent Bayesian classifier to determine the likelihood ratio, indicating whether gene pairs within each bin contained an overrepresentation of truly interacting gene pairs. Once the likelihood ratios

had been determined for each data set, a receiver operator characteristic (ROC) curve was constructed, and the area under the curve (AUC) was calculated. Data sets that had a minimal AUC of 0.59 were combined in a naive way—for each gene pair, the likelihood ratios were multiplied by each other, resulting in a final microarray coexpression likelihood ratio for each gene pair.

Two orthologous protein-protein interaction data sets from Lehner and Fraser[20] were used to supplement the GO and microarray coexpression data. One data set contained computationally predicted human protein interactions that had been physically mapped within Ensembl genes. The second data set contained a subset of these protein pairs, to which Lehner *et al*. had assigned a higher confidence. Three bins were constructed: one containing the higher-confidence gene pairs, one containing the remaining lower-confidence pairs, and a third containing all the other unobserved gene pairs. A human Y2H protein-protein interaction data set from Stelzl *et al*.[19] was integrated by mapping the HUGO identifiers to Ensembl genes. Two bins were constructed: one containing the gene pairs for which a Y2H interaction was reported, and one containing all the other unobserved gene pairs.

Network Integration

The Bayesian classifier was employed to integrate the various binned types of data. We chose not to learn the Bayesian network structure from the data but to use a predefined Bayesian network structure, for which the conditional probabilities were determined by benchmarking the various data sets against the gold standard (fig. 2) (details provided in appendix A). We subsequently generated four gene networks. One network contained evidence for interaction based on the GO data (GO network). Another network contained evidence for interaction derived from integrating the microarray coexpression and predicted protein-protein interaction data in a naive way (MA+PPI network). A third network combined, in a naive way, the GO and MA+PPI networks (GO+MA+PPI network), and this was complemented with all known true-positive interactions in a final network (GO+MA+PPI+TP network). To relate interacting genes directly or indirectly, an all-pairs shortest path was calculated for each gene network[35]. This measure of theminimal path length between pairs of genes was used in the subsequent method to associate disease genes with each other.

Disease Analysis and Positional Candidate-Gene Prioritization

Prioritizer assesses whether genes residing within different susceptibility loci are close together within the gene network. This indicates that this method could also work with diseases for which only two loci have been identified. However, in such a case, there is a considerable probability that two genes, each residing in a different locus, would interact by chance. We therefore restricted the analysis to diseases for which at least three contributing disease genes had been identified. These diseases and disease genes were derived from the Online Mendelian Inheritance in Man (OMIM) database[36], by text mining the first paragraphs of all OMIM disease entries as of March 1, 2005, and extracting the OMIM gene numbers contained within these paragraphs. The HUGO gene name was later extracted from these OMIM entries and was mapped to an Ensembl gene name. If, for any one disease, there were two disease genes situated at the same chromosome and positionally less than 200 genes apart, one of the two genes was randomly removed to ensure that no loci would overlap.

The diseases for which at least three disease genes remained after filtering were analyzed by artificially generating susceptibility loci around the disease genes, in a range from 50 to 150 genes, in steps of 50. All 20,334 genes were assigned an initial effect score of zero, and, subsequently, all loci were traversed. Using each gene network for all positional candidate genes residing in a particular locus, we determined whether any of these genes were functionally closely related to genes physically residing inside another susceptibility locus. If this was the case, the effect score of the related gene that was functionally close but physically in another locus increased (fig. 1), by use of the following Gaussian kernel scoring function:

$$\text{effect} = e^{-\left(\frac{\text{distance}^2}{53}\right)}$$

Figure 2

Integration of data sets in four gene networks. **a)** Data sets were benchmarked against a set of 55,606 known true-positive gene pairs derived from BIND, KEGG, HPRD, and Reactome and 800,608 true-negative gene pairs derived from GO. The Venn diagram indicates the data sources from which the true positives were derived and their degree of overlap. Numbers in parentheses indicate the number of interactions that are provided by each of the data sets. **b)** Potential gene-gene interactions derived from GO, microarray coexpression data, and human and orthologous protein-protein interaction data were integrated using a Bayesian classifier. The steps involved in building this classifier are shown.

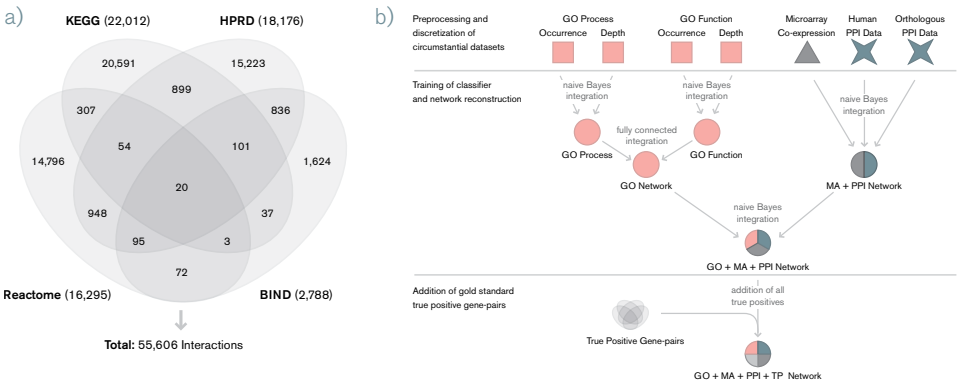
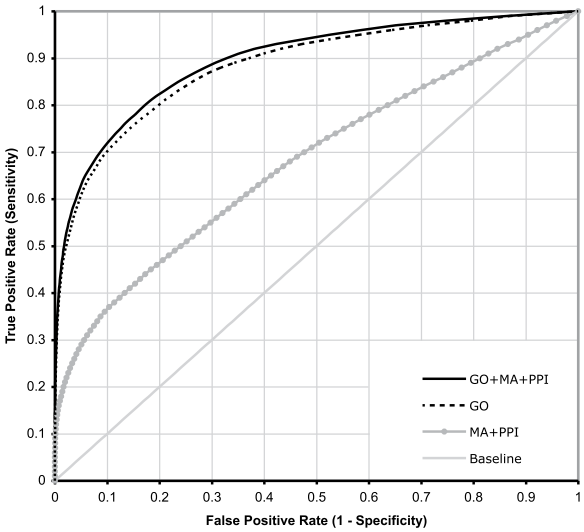


Figure 3

ROC curve of the GO network, the MA+PPI network, and the combined GO+MA+PPI network. The baseline (solid gray line) indicates the performance of a classifier that would be totally uninformative.



where "distance" is defined as the all-pairs shortest path between the two genes. The kernel function width was chosen arbitrarily, but a sensitivity analysis showed that different widths did not influence the results much (data not shown). By applying this function, positional candidate genes that resided in different loci but that were functionally closely related in the gene network were assigned higher scores than positional candidate genes that were functionally far apart from each other. To correct for differences in topology of the gene network, an empiric P value was determined for each positional candidate gene through permutation of the other loci 500 times by reshuffling them across the genome and recalculating the effect scores. This permitted a probability density function to be determined per positional candidate gene, for which the empiric P value could be looked up. For each locus, the positional candidate genes were prioritized on the basis of this P value.

Results

Construction of a Functional Gene Network

The basis for our human gene network was a gold standard of validated gene-gene interactions (true positives) and a further set of gene-gene pairs that were deemed highly unlikely to interact (true negatives). To construct the set of true-positive gene pairs, 2,788 confirmed, direct, physical protein-protein interactions were derived from BIND; 18,176 confirmed human protein interactions were derived from HPRD; 22,012 direct functional interactions were derived from KEGG; and 16,295 interactions were derived from Reactome. This resulted in 55,606 unique true-positive gene relationships (fig. 2a). For the true-negative set, gene pairs were selected that encode for proteins localized in different cellular compartments. The combinations of cellular compartments were selected from their underrepresentation in the set of true positives (see the "Material and Methods" section). This resulted in 801,108 pairs, of which 500 were known to be true-positive gene pairs, and these were therefore removed from the set of true-negative gene pairs.

We trained the classifier on this gold standard and constructed functional human gene networks on the basis of GO data, microarray coexpression data, and inferred protein-protein interactions, as well as combinations of these. First, for each gene pair, we assessed whether the genes shared GO annotations, which were derived for 15,045 genes from Ensembl. Sharing of GO terms was based on the frequency of the least-common GO term shared between two genes and the maximal depth in the GO hierarchy at which two shared terms lay. Gene coexpression was calculated in 186 microarray data sets derived from GEO and 75 data sets from SMD. However, most of these data sets were not highly informative, as judged by their ability to identify true-positive gene interactions with a low false-positive rate. Because it is known that many classifiers perform best when a subset of features are used[37,38], we used only four informative microarray coexpression data sets for classification[39–42], each showing a minimal AUC of 0.59. In total, these data sets contained 461 microarray hybridizations. Finally, protein-protein interactions were derived from the Lehner and Fraser[20] data set containing human protein interactions predicted by mapping physical protein interactions from various *Saccharomyces cerevisiae*, *Drosophila melanogaster*, and *Caenorhabditis elegans* interaction data sets to orthologous human gene pairs. Of the 71,806 predicted gene pairs, we were able to physically map 62,635 gene pairs with both genes in the pair mapping to known Ensembl genes. A subset was defined by Lehner and Fraser[20] that contained 10,652 gene pairs deemed to be of higher confidence, of which 10,139 gene pairs could be mapped. In addition, we used 3,186 human protein-protein interactions identified by automated Y2H interaction mating by Stelzl *et al*[19], of which 1,751 could be mapped to different Ensembl gene pairs.

We assessed the performance of our classifier on the basis of these various data sources in three different gene networks generated on the basis of a Bayesian framework, after preprocessing and binning of the data sets. As mentioned above, one network was generated solely on the basis of GO

data (GO network), one network was based on both microarray coexpression and predicted protein-protein interaction data (MA+PPI network), and an overall network contained all three types of data (GO+MA+PPI network). ROC curves (fig. 3) show the performance of the reconstructed GO,MA+PPI, and GO+MA+PPI gene networks, which were constructed by cross-validating all data sets 10 times against the gold standard set, to mitigate overfitting (details provided in appendix A). When we compared the performance of the various gene networks, it became evident that the GO data set provided the most accurate evidence for interaction. The AUC was 88%, compared with 50% for an uninformative classifier. The ROC for the MA+PPI network shows that coexpression data derived from microarray expression, in conjunction with the orthologous protein-protein interaction data, correctly inferred functional interactions (AUC = 68%), but to a lesser extent than the GO network. Nevertheless, as can be deduced from the GO+MA+PPI network, addition of the microarray coexpression and the orthologous protein- protein interaction data to the GO network improved slightly the accuracy of the network (AUC = 90%). In accordance with most networks described in the literature thus far[43], our reconstructed networks have a connectivity that follows a scale-free power-law distribution, which has also been demonstrated for other organisms [44–46].This is most apparent when the topology of the MA+PPI network is assessed (see appendix A). To validate our network, we used a list of 2,574 Y2H interactions that recently became available47 to assess whether our gene network had predicted an interaction for these gene pairs. We first mapped the set to Ensembl pairs and then removed all pairs that were in our gold standard true-positive set, to ensure that we only assessed newly identified interactions. This resulted in a set of 1,318 novel gene pairs.

We then assessed whether our gene network had predicted an interaction for these pairs. While Y2H interactions are known to regularly yield false-positive results[48], we decided to test whether the distribution of likelihood ratios for these gene pairs was significantly different from a null distribution of 10,000 gene pairs sampled by generation of random gene pairs by selecting two genes at a time from the set of all individual genes that made up the Y2H gene pairs. The results show that the 1,318 Y2H gene pairs have a significantly higher likelihood ratio than the null distribution (P = .0003, by Wilcoxon Mann-Whitney test), which indicates that our gene network is capable of inferring as-yet-unknown interactions.

To allow researchers to look up known and predicted interactions and to identify the shortest routes between genes and susceptibility loci, we developed a Web tool, which is publicly available at the GeneNetwork-Web site. The known and predicted interactions can be shown for each gene of interest, along with information about the source of evidence from which they were derived and how strong this evidence was. In addition, there are interactive graphs to visually explore how multiple genes interact with each other. All the data files (including the sets of true-positive and true-negative gene pairs) can be downloaded, along with a Java application programming interface, which can facilitate the development of new methods that use this gene network. We will regularly update the gene network, on the basis of the most recent releases of the various repositories used in its construction.

Increased Functional Interactions Shown by Genes Associated with a Particular Disease

We first examined our hypothesis that genes associated with genetic disorders frequently share functional links, by assessing whether, for a disease, these causative genes were functionally more closely related to each other than a set of genes of equal size that were randomly selected from the full set of 345 unique disease genes of the 409 disease genes that were extracted from OMIM entries on disorders for which at least three causative genes were known. This set of disease genes was used as a background distribution to prevent bias, since the disease genes are generally better characterized than the complete set of genes in the network. We generated one extra network (GO+MA+PPI+TP network) that complemented the GO+MA+PPI network with all known true-positive gene pairs,

and we calculated the shortest direct or indirect distance between all pairs of genes. In 76 (79%) of the 96 diseases, the total distance between all combinations of disease genes in one disease was, on average, lower than the total distance between all combinations of randomly selected disease genes in 10,000 permutations. This confirms our hypothesis that, in the majority of diseases, the causative genes are indeed closely related functionally.

Genes implicated in disease processes tend to be studied more than those not implicated, which could result in a bias in the gene network based on GO annotations, since these represent known functional annotations. To assess the degree to which this possible bias affected our gene network, we looked at network connectivity. The average number of direct interactions involving disease genes was 199, compared with an average of 203 for the other 11,875 genes that interacted with at least one other gene. This indicates that other genes are equally represented in the gene network, despite the fact that disease genes may have been studied more.

Increased Power to Detect Disease Genes Provided by a Functional Gene Network

Usually, researchers pick a limited number of candidate genes in susceptibility loci to follow-up, because it is too costly and labor intensive to analyze all the genes residing in these loci. As a result, these studies have a limited chance of finding disease-related variants, largely depending on the size of the loci and the number of genes selected. Using a test set of known disorders in a similar setup, we evaluated the ability of our reconstructed network to correctly prioritize positional candidate genes in a set of top-ranked candidate genes of typical size (5–10 genes). The test set consisted of 96 different disorders, for which a total of 409 disease genes (345 unique genes) had been identified. These were obtained from OMIM, with 3–10 disease genes per disease (average 4.3 genes per disease). Of the diseases, 59 are of Mendelian origin, 17 have complex inheritance, and 20 are various types of cancer (table 1).

The ability of the functional human gene network to correctly prioritize known disease genes was assessed by creating artificial, nonoverlapping susceptibility loci around these disease genes. Since many genes in these loci have no known or predicted interactions in our network, we only assessed those genes for which interactions were predicted, to prevent a bias toward genes that were better represented in the underlying high-throughput data sets. This resulted in susceptibility loci of varying widths, containing 50, 100, or 150 genes, which were predicted to interact with at least one other gene. If, for any particular disease, two disease genes residing in the same chromosome yielded loci that were partly overlapping, one of the two loci was randomly removed. For each locus, the genes were traversed, and, for each gene, we assessed whether there was another gene residing in a different locus that was nearby within the gene network. The effect scores (see the "Material and Methods" section) of each gene were affected by the gene in the other locus that had the shortest path to that gene. This procedure has the potential to preferentially identify genes with many interacting partners over genes that are less well connected, because a highly connected gene has a higher chance of interacting with a gene residing in another locus than a gene for which only a few interactions have been predicted. To overcome bias in the method toward genes that are highly connected, we corrected for differences in the network topology by permuting the susceptibility loci for each disease 500 times across the genome.

After all positional candidate genes were ranked on the basis of this permuted score, the results (see fig. 4 and table 1) indicated that this method was able to identify many of the disease genes in the top 5 or top 10 genes per locus. As expected from the ROC curves of the various gene networks (fig. 3), the performance of the MA+PPI network proved to be the least powerful. Nevertheless, the number of correctly ranked genes was higher than would be expected to occur by chance (fig. 4a and 4b; indicated by baseline) for many of the susceptibility loci widths. When assessing susceptibility loci that contained, on average,

Table 1

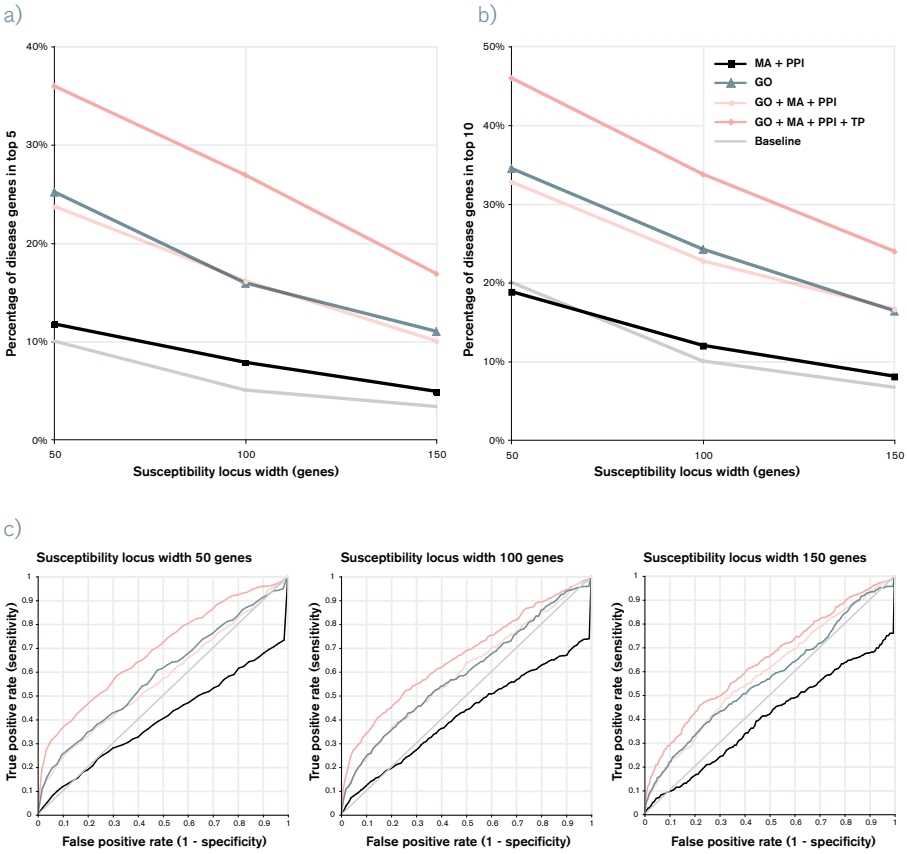
Overview of the 96 diseases studied with Prioritizer and the number of disease genes per disorder that ranked in the top 10 genes per susceptibility locus, with locus widths of 100 and 150 genes

	Diseases	OMIM number	No of genes	Number of genes ranking in top-10 at locus width 100 genes				Number of genes ranking in top-10 at locus width 150 genes			
				MA+PPI	GO	GO+MA+PPI	GO+MA+PPI+TP	MA+PPI	GO	GO+MA+PPI	GO+MA+PPI+TP
Mendelian inheritance	Achromatopsia 2	216900	3	0	0	1	0	0	0	0	0
	Achromatopsia 3	262300	3	0	0	1	0	0	0	0	0
	Adrenoleukodystrophy, autosomal neonatal form	202370	5	0	4	2	4	0	4	2	3
	Amyloidosis VI	105150	3	1	1	0	1	0	0	0	1
	Amyloidosis, familial visceral	105200	3	0	0	0	0	0	0	0	0
	Amyotrophic lateral sclerosis 1	105400	5	0	1	0	0	0	1	0	0
	Atypical mycobacteriosis, familial	209950	5	1	4	2	4	1	1	0	1
	Autonomic control, congenital failure of	209880	5	0	0	1	1	0	0	1	1
	Bardet-biedl syndrome	209900	8	0	2	2	3	0	1	1	1
	Bare lymphocyte syndrome, type II	209920	4	1	0	1	4	0	0	1	4
	Cardiomyopathy, familial hypertrophic	192600	9	0	7	4	4	0	7	3	3
	Cholestasis, intrahepatic, of pregnancy	147480	3	1	0	0	0	0	0	0	0
	Cholestasis, progressive familial intrahepatic 1	211600	4	1	1	1	0	0	1	1	1
	Complex I, mitochondrial respiratory chain, deficiency of	252010	5	0	5	4	5	0	3	3	3
	Coumarin resistance	122700	4	0	0	0	1	0	0	0	0
	Dementia, lewy body	127750	3	1	0	0	0	0	0	0	0
	Epidermolysis bullosa junctionalis, disentis type	226650	4	1	0	1	0	0	0	0	0
	Epidermolysis bullosa of hands and feet	131800	4	0	1	2	1	0	0	1	1
	Fanconi anemia	227650	6	1	0	1	6	1	0	0	6
	Fundus albipunctatus	136880	4	0	1	0	3	0	1	1	2
	Generalized epilepsy with febrile seizures plus	604233	3	2	2	2	1	1	0	1	0
	Glutaricaciduria iia	231680	3	3	2	3	3	3	1	3	3
	Hermansky-pudlak syndrome	203300	6	0	1	0	0	0	0	0	0
	Hirschsprung disease	142623	6	1	0	0	1	1	0	0	1
	Hydrops fetalis, idiopathic	236750	4	0	0	1	0	0	0	1	0
	Hypercholesterolemia, familial	143890	6	0	2	3	1	0	1	2	1
	Hypertrophic neuropathy of dejerine-sottas	145900	4	0	2	2	2	1	1	1	1
	Hypokalemic periodic paralysis	170400	3	0	0	0	2	0	0	0	0
	Ichthyosiform erythroderma, congenital, nonbullous, 1	242100	3	0	2	2	1	0	1	1	1
	Immunodeficiency with hyper-IGM, type 2	605258	3	0	1	0	2	0	1	0	1
	Immunodeficiency with hyper-IGM, type 3	606843	3	0	2	0	2	0	1	0	1
	Kartagener syndrome	244400	3	1	2	2	1	0	2	2	1
	Keratosis palmoplantaris striata I	148700	3	0	1	1	3	0	1	1	3
	Laron syndrome, type II	245590	3	0	1	0	2	0	0	0	1
	Leber congenital amaurosis, type I	204000	7	0	3	2	1	0	0	1	2
	Leigh syndrome	256000	6	3	4	3	3	1	4	3	3
	Leukoencephalopathy with vanishing white matter	603896	5	5	5	5	5	5	5	5	4
	Maple syrup urine disease, type Ia	248600	4	1	1	1	4	0	0	0	3
	Maturity-onset diabetes of the young	606391	5	1	0	1	3	1	0	1	0
	Myasthenic syndrome, congenital, fast-channel	608930	3	0	2	2	3	0	2	2	3
	Myasthenic syndrome, slow-channel congenital	601462	3	0	2	2	1	0	2	2	2
	Myoclonic dystonia	159900	3	0	0	0	1	0	0	0	1
	Nemaline myopathy 1, autosomal dominant	161800	3	0	1	1	0	0	1	1	0

	Nesidioblastosis of pancreas	256450	3	0	1	0	0	0	1	1	0
	Night blindness, congenital stationary	163500	3	0	0	1	0	0	0	0	0
	Obsessive-compulsive disorder 1	164230	3	0	1	0	0	0	0	0	0
	Ossification of posterior longitudinal ligament of spine	602475	3	0	0	0	1	0	0	0	0
	Osteopetrosis, autosomal recessive	259700	3	1	0	0	0	0	0	0	0
	Peters anomaly	604229	4	0	0	1	2	0	0	0	0
	Pituitary dwarfism III	262600	3	0	0	0	2	0	0	0	1
	Progressive external ophthalmoplegia	157640	3	1	1	0	0	0	0	0	0
	Pseudohypoaldosteronism, type I, autosomal recessive	264350	3	0	2	2	1	0	2	2	0
	Pulmonary alveolar proteinosis	265120	3	0	2	1	0	0	2	1	0
	Refsum disease, infantile form	266510	3	1	1	1	2	0	1	1	2
	Reticulosis, familial histiocytic	267700	3	0	0	0	0	0	0	0	0
	Rhizomelic chondrodysplasia punctata, type 3	600121	3	0	1	0	2	0	1	0	1
	Stickler syndrome, type I	108300	3	0	0	0	2	0	0	0	0
	Waardenburg-shah syndrome	277580	3	0	1	1	1	0	2	1	0
	Zellweger syndrome	214100	8	1	4	4	7	1	3	4	5
Complex inheritance	Alzheimer disease	104300	8	0	1	0	3	0	1	1	2
	Diabetes mellitus, noninsulin-dependent	125853	9	2	0	3	1	1	0	2	2
	Elliptocytosis, rhesus-unlinked type	130600	3	0	1	0	3	0	1	0	2
	Graves disease	275000	3	0	1	0	1	0	0	0	0
	Hypertension, essential	145500	7	1	1	0	0	0	0	0	0
	Hypospadias	146450	3	0	0	0	1	0	0	0	1
	Iga nephropathy	161950	4	1	0	0	1	1	0	0	1
	Inflammatory bowel disease 1	266600	4	0	0	1	1	0	0	0	1
	Longevity	152430	4	0	1	0	0	1	0	0	0
	Lupus erythematosus, systemic	152700	4	0	0	0	0	0	0	0	0
	Mycobacterium tuberculosis, susceptibility to infection	607948	3	0	0	0	0	0	0	0	0
	Myoclonic epilepsy, juvenile	606904	4	0	1	0	1	0	1	0	0
	Obesity	601665	7	1	1	1	4	2	0	1	3
	Osteoporosis, involutional	166710	5	0	1	1	0	0	3	1	2
	Parkinson disease	168600	4	0	0	0	4	0	1	0	3
	Rheumatoid arthritis	180300	5	0	0	0	0	0	0	0	1
	Sudden infant death syndrome	272120	3	0	2	2	0	0	1	1	0
	Bladder cancer	109800	3	0	0	0	0	0	0	1	0
Heritable cancer	Breast cancer	114480	10	2	1	4	2	1	0	2	1
	Chondrosarcoma	215300	4	1	1	0	2	0	0	0	1
	Esophageal cancer	133239	8	1	0	1	5	1	0	0	2
	Glioma of brain, familial	137800	6	1	1	1	0	2	1	0	0
	Hepatocellular carcinoma	114550	3	0	0	0	1	1	0	0	0
	Juvenile myelomonocytic leukemia	607785	4	0	3	2	1	0	1	1	1
	Leiomyoma, uterine	150699	4	0	0	1	0	0	0	0	0
	Lung cancer	211980	4	1	0	1	2	0	0	1	0
	Lymphoma, non-hodgkin, familial	605027	4	0	2	2	2	0	1	1	2
	Medulloblastoma	155255	4	1	0	2	2	1	0	1	0
	Myeloma, multiple	254500	4	1	1	0	0	1	0	0	1
	Osteogenic sarcoma	259500	3	0	0	0	0	0	0	0	1
	Pancreatic carcinoma	260350	6	1	1	1	0	1	0	1	0
	Pheochromocytoma	171300	3	0	0	0	0	0	0	0	0
	Prostate cancer	176807	9	1	0	1	0	0	0	0	0
	Renal cell carcinoma, papillary	605074	3	1	0	0	1	1	0	0	1
	Rhabdomyosarcoma 2	268220	3	0	0	0	0	0	0	0	0
	Thyroid carcinoma, papillary	188550	5	2	0	0	0	1	0	0	0
	Turcot syndrome	276300	3	2	2	2	1	2	2	3	2
		Total	**409**	**49**	**99**	**93**	**138**	**33**	**67**	**68**	**98**

Figure 4

Accuracy of positional candidate-gene prioritization. **a)** and **b)**, Percentage of the 409 disease genes that was ranked among the top 5 (a) or top 10 (b) genes per locus, after artificial susceptibility loci of varying widths around these genes were constructed and when different types of gene networks were used. The baselines (gray lines) indicate the percentage of disease genes expected to rank among the top 5 or top 10 genes by chance. **c)** ROC curves for susceptibility loci that contain 50, 100, or 150 genes.



100 genes, we found 8% and 12% of the disease genes were contained within the top 5 and top 10 per locus, respectively, compared with the 5% and 10% we would expect to find by chance. A lack of predictive performance of the MA+PPI network explains why the ranking did not improve considerably when this network was used, as is evident from inspection of the ROC curves (fig. 4c), which show the proportion of disease genes and nondisease genes that are returned when different sizes of sets of top-ranked genes per locus are assessed. For 86 of the 345 unique disease genes within the MA+PPI network, no interactions were predicted. Hence, they were ranked low, the more so because the 49, 99, or 149 other genes, residing together with each disease gene in the constructed susceptibility loci, had been selected on the premise that they interacted with at least one other gene. The GO network performed considerably better; when we used it to assess susceptibility loci that contained, on average, 100 genes, we found 16% and 24% of the disease genes were contained within the top 5 and top 10 genes per locus, respectively. The performance of the disease analysis was best when the inferred GO+MA+PPI network was complemented with the known true-positive interactions (GO+MA+PPI+TP network); with this network and an average susceptibility locus width of 100 genes, 27% and 34% of the disease genes were contained within the top 5 and top 10 per locus, respectively.

We also assessed the probability of detecting at least one disease gene when only a fixed number of topranked genes per locus is followed up (fig. 5). When we employed the most comprehensive GO+MA+PPI+TP network and followed up all the top 5 or top 10 positional candidate genes for each disorder, using locus widths of 100 and 150 genes, we found at least one disease gene from these top sets of genes in 54% and 64% of the diseases, respectively, compared with 19% and 35% expected by chance. When we confined our analysis to diseases for which at least four or five disease genes were known, the performance of our method increased slightly (data not shown), because the true disease genes now interacted with more of the other true disease genes, increasing their overall scores.

Breast Cancer as an Example

We selected breast cancer as an example of how the various gene networks perform in a complex disease for which multiple disease genes have been identified. Artificial susceptibility loci, each comprising 100 genes, were constructed around 10 putative breast cancer genes described in OMIM (as of March 1, 2005). For each of the four networks, we then determined how many of the disease genes were ranked within the top 10 per locus. The MA+PPI network ranked two disease genes (*PIK3CA* and *CHEK2*) in the top 10, whereas the GO network ranked three (*BRCA2*, *NCOA3*, and *CHEK2*), and the GO+MA+PPI network ranked four (*BARD1*, *PIK3CA*, *TP53*, and *CHEK*) (fig. 6). However, the GO+MA+PPI+TP network, which integrates the most information, performed the worst; of the 10 disease genes now known, only 2 (*BARD1* and *BRCA1*) were ranked in the top 10. This can be explained by the observation that the true-positive set contained many known interactions for these 10 breast cancer genes. As the ranking procedure corrects for the topology of the network, these disease genes, with a marked increase in the number of relationships with other genes in this most comprehensive network, were suddenly no longer ranked as high. This became evident when the genes were ranked using the GO+MA+PPI+TP network but the differences in topology were not corrected for: 9 of the 10 breast cancer genes were then in the top 10 per locus.

Prioritizer Availability

To allow researchers to analyze susceptibility loci of interest, we developed a Java application that can be downloaded, along with regularly updated gene network definition files and source code from the Prioritizer Website. After a set of susceptibility loci has been entered, Prioritizer ranks the positional candidate genes in each locus by using the method described above in conjunction with one of the four gene networks. It can generate two- and three-dimensional graphs of the top-ranked positional candidate genes, which allows the user to visually inspect how the genes within the different loci interact with each other.

Figure 5

Probability of detecting at least one disease gene when a fixed number of top-ranked positional candidate genes—as ranked by Prioritizer—are followed up for each locus. Each locus contains either 100 or 150 genes, and the GO+MA+PPI+TP network was employed.The baselines (dashed lines) show the probability of detecting at least one disease gene if a fixed number of arbitrarily chosen genes in each locus are followed up.

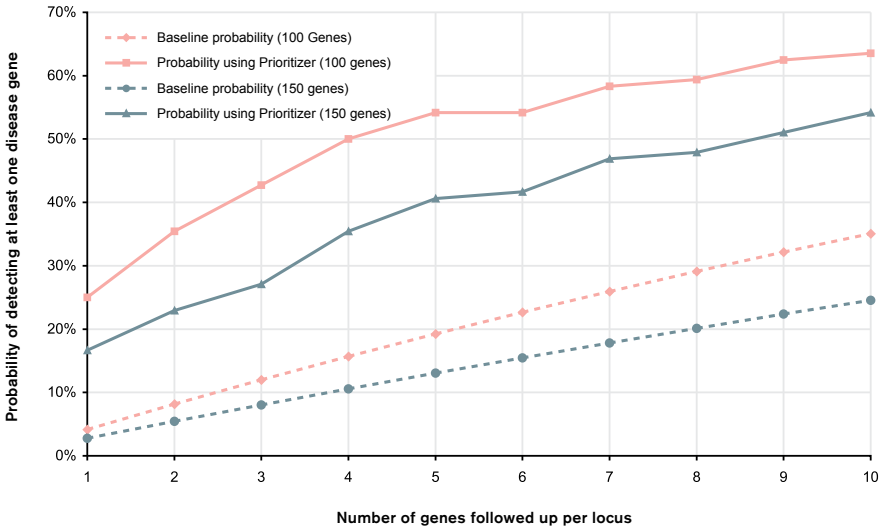
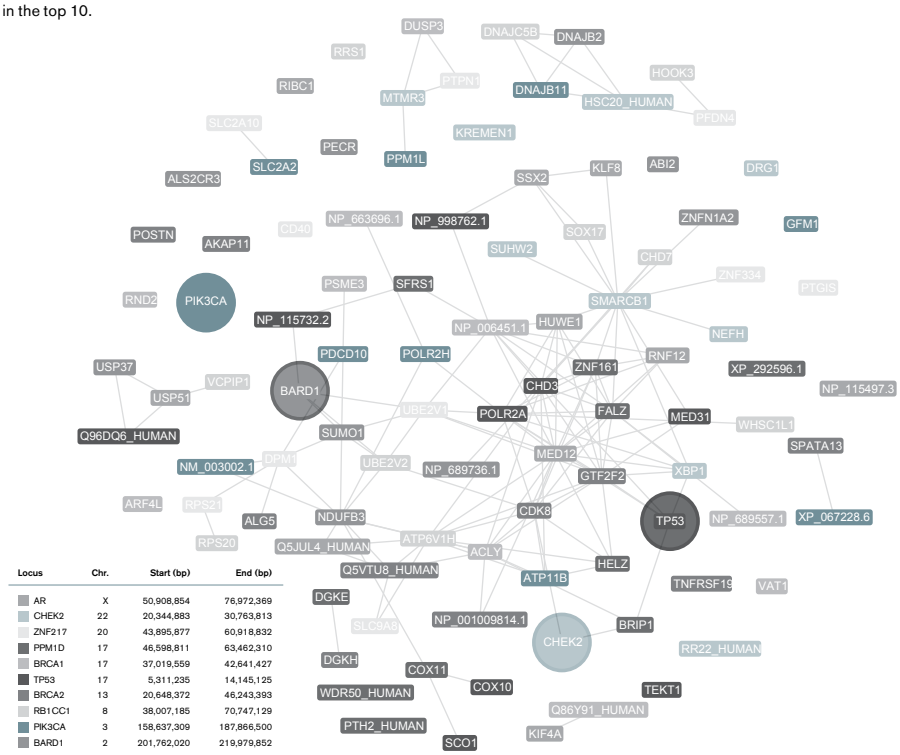


Figure 6

Prioritizer analysis of breast cancer. Susceptibility loci, each containing 100 genes, were defined around 10 known breast cancer genes. The 10 highest-ranked genes for each locus are shown in the graph, with colors indicating the locus in which they reside. Use of the GO+MA+PPI network led to four breast cancer genes (PIK3CA, CHEK2, BARD1, and TP53 [*circles*]) being ranked in the top 10.



Discussion

In this study, we describe the construction of a functional human gene network of considerable accuracy (fig. 3; AUC = 90%). As such, it can be used to assess interactions for a gene of interest through the bioinformatics tools that we have made available online. We have shown that, in cases where multiple genes underlie a disorder, these genes tend to have more functional interactions. When these functional interactions are employed to prioritize known disease genes in artificial susceptibility loci, the chance of detecting disease genes is increased considerably (2.8 fold).

In breast cancer, 4 of the 10 disease genes were ranked in the top 10 when the GO+MA+PPI network was applied, a fourfold enrichment over the single disease gene that would be picked up by chance. As has been discussed earlier, the correction for differences in topology is needed to prevent bias toward highly connected genes. However, this puts diseases in which underlying genes have a high degree of connectivity at a disadvantage, which was apparent in the analysis of breast cancer by use of the GO+MA+PPI+TP network. When this topology correction was omitted for breast cancer, the ranking of the disease genes improved considerably, to include 9 of the 10 genes. The availability of new highthroughput data sets will alleviate this problem in the future, by providing novel interactions for genes that currently have a low degree of connectivity, which will reduce the penalty on highly connected genes.

We noticed that the performance of Prioritizer was lower for complex disorders than for Mendelian disorders. This is likely caused by the fact that the etiology of complex diseases is more subtle and involves multiple pathways, so that most of the disease genes only confer a modest increased risk. Greater coverage of the gene network, leading to identification of relationships between genes that bridge the various pathways, could probably help to alleviate this problem.

When the accuracy of the various gene networks was assessed by investigation of their respective ROCs, it was envisaged that the GO+MA+PPI network would perform at least at a similar level in prioritizing disease genes as the GO network, because its AUC was greater. However, contrary to our expectation, when the positional candidate genes were prioritized, the disease genes in some diseases were ranked lower with the GO+MA+PPI network than with the GO network. One explanation could be that, within the microarray coexpression data sets (the main contributor to the MA+PPI network), we did not distinguish between coexpression and coregulation. As such, many direct interactions between genes were inferred, but a large proportion of these interactions were actually indirect. Methods have recently appeared[33,49] that could help remove some of these incorrectly inferred interactions.

In a somewhat comparable method by Turner *et al*[21], positional candidate genes are prioritized by determining which genes share InterPro[50] domains and GO terms, as a measure to relate genes in susceptibility loci with each other. Our method extends this approach by also allowing for indirect relationships between individual disease genes, since Prioritizer uses the graph-theoretic distance between genes to relate them. Both approaches still rely largely on manual annotation, which is detrimental for genes that have not been investigated extensively. When no experimental evidence for interaction is available, there is only a small chance that these potential disease genes, residing in one specific susceptibility locus, will be associated with disease genes in other loci, since the sharing of GO or InterPro terms between these genes will be minimal. Although GO contributes the most to the performance of the Bayesian classifier, we should not depend entirely on a prediction if there is substantial evidence only from GO, while the evidence from the other data sets is lacking, for a specific gene pair, because the GO evidence has been inferred from the sharing of predominantly manually annotated terms, whereas the other sources rely more on direct biological measurements. It is expected that, when additional high-throughput data sets become available and

their coverage of all possible functional interactions increases, GO evidence will be supplemented by experimental data, resulting in better predictions.

As such, an extensive and reliable functional gene network is crucial for good performance of our method. If this network is inaccurate or biased toward known genes, the ranking of true disease genes in the susceptibility loci will deteriorate. Several rapidly expanding data repositories are now becoming available that should help to improve our network. They include text mining methods[51,52], which extract functional relationships from the literature, and methods that integrate results from high-throughput proteomic approaches.53 Our gene network, which, in its current form, has been applied to genetic linkage analysis, can also be used for other applications. Recently, efforts have been made to prioritize positional candidate genes on the basis of their expression[54], with the assumption that differences in expression behavior in comparisons of patients with controls may be due to cis-acting variants in the underlying genes. However, it has turned out that, in most genes, differences in expression are determined by genetic variation in genes located elsewhere[55,56]. The reconstructed functional gene network can help to relate the observed differences in gene expression to the underlying causative genetic variants in other genes, which might help in identifying the disease genes.

Prioritizer might also be well suited for genomewide SNP association studies. Technical improvements in conjunction with decreasing costs now allow researchers to perform these studies in complex diseases, thereby considerably increasing the resolution at which one can assess genetic variation. However, as the number of tested SNPs increases, the number of tested individuals required to achieve sufficient power will also rise. To help overcome this problem, a new statistical method has recently been developed[57] that combines evidence from the most-significant tests, under the assumption that there are multiple true associations in the disease under investigation. However, within this confined set, the majority of genes will still be false positives because of power issues. Our positional candidate-gene prioritization method can easily be adapted to help distinguish true disease-associated genes and false-positive genes, by assuming that the true disease genes are mostly functionally related and will therefore be closer to each other in the gene network than to the false-positive genes that have been randomly selected.

We have demonstrated that it is feasible to use gene networks to prioritize positional candidate genes in various heritable disorders with multiple associated genes, even when the susceptibility loci are fairly large. As such, this article and the proposed methods show that the integration of gene networks with various genetic studies can be useful in identifying disease genes. We envisage that improvements both in the quality of the data sets making up these gene networks and in the statistical methods incorporating the networks will result in new, genetically testable hypotheses.

Acknowledgments

We thank Jackie Senior and members of the Complex Genetics Section and the Department of Human Genetics for critically reading the manuscript. This study was supported by Netherlands Organization for Scientific Research grant 901-04-219 and by a grant from the Celiac Disease Consortium, an innovative cluster approved by the Netherlands Genomics Initiative and partially funded by a Dutch government grant (BSIK03009).

Web Resources

The URLs for data presented herein are as follows:

Biomolecular Interaction Network Database (BIND), bind.ca

Ensembl, www.ensembl.org

GeneNetwork, www.genenetwork.nl

Human Protein Reference Database (HPRD), www.hprd.org

Kyoto Encyclopedia of Genes and Genomes (KEGG), www.genome.jp/kegg

Online Mendelian Inheritance in Man (OMIM), www.ncbi.nlm.nih.gov/Omim

Prioritizer, www.prioritizer.nl

Reactome, www.reactome.org

References

1 Jacobi FK, Broghammer M, Pesch K, Zrenner E, BergerW, Meindl A, Pusch CM (2000) Physical mapping and exclusion of GPR34 as the causative gene for congenital stationary night blindness type 1. Hum Genet 107:89–91

2 Seri M, Martucciello G, Paleari L, Bolino A, Priolo M, Salemi G, Forabosco P, Caroli F, Cusano R, Tocco T, Lerone M, Cama A, Torre M, Guys JM, Romeo G, Jasonni V (1999) Exclusion of the Sonic Hedgehog gene as responsible for Currarino syndrome and anorectal malformations with sacral hypodevelopment. Hum Genet 104:108–110

3 Simard J, Feunteun J, Lenoir G, Tonin P, Normand T, Luu The V, Vivier A, *et al* (1993) Genetic mapping of the breast-ovarian cancer syndrome to a small interval on chromosome 17q12-21: exclusion of candidate genes EDH17B2 and RARA. Hum Mol Genet 2:1193–1199

4 Tumer Z, Croucher PJ, Jensen LR, Hampe J, Hansen C, Kalscheuer V, Ropers HH, Tommerup N, Schreiber S (2002) Genomic structure, chromosome mapping and expression analysis of the human AVIL gene, and its exclusion as a candidate for locus for inflammatory bowel disease at 12q13-14 (IBD2). Gene 288:179– 185

5 Walpole SM, Ronce N, Grayson C, Dessay B, Yates JR, Trump D, Toutain A (1999) Exclusion of RAI2 as the causative gene for Nance-Horan syndrome. Hum Genet 104:410–411

6 Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, *et al* (1994) A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science 266:66-71

7 Joenje H, Patel KJ (2001) The emerging genetic and molecular basis of Fanconi anaemia. Nat Rev Genet 2:446–457

8 D'Andrea AD, Grompe M (2003) The Fanconi anaemia/BRCA pathway. Nat Rev Cancer 3:23–34

9 de Winter JP, van der Weel L, de Groot J, Stone S, Waisfisz Q, Arwert F, Scheper RJ, Kruyt FA, Hoatlin ME, Joenje H (2000) The Fanconi anemia protein FANCF forms a nuclear complex with FANCA, FANCC and FANCG. Hum Mol Genet 9:2665–2674

10 Yamashita T, Kupfer GM, Naf D, Suliman A, Joenje H, Asano S, D'Andrea AD (1998) The Fanconi anemia pathway requires FAA phosphorylation and FAA/FAC nuclear accumulation. Proc Natl Acad Sci USA 95:13085–13090

11 Zatz M, de Paula F, Starling A, Vainzof M (2003) The 10 autosomal recessive limb-girdle muscular dystrophies. Neuromuscul Disord 13:532–544

12 Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, *et al* (2005) The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res Database Issue 33:D418–D424

13 Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, Gandhi TK, *et al* (2004) Human Protein Reference Database as a discovery resource for proteomics. Nucleic Acids Res Database Issue 32:D497–D501

14 Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res Database Issue 32:D277–D280
15 Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L (2005) Reactome: a knowledgebase of biological pathways. Nucleic Acids Res Database Issue 33:D428– D432
16 Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, *et al* (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res Database Issue 32:D258– D261
17 Ball CA, Awad IA, Demeter J, Gollub J, Hebert JM, Hernandez- Boussard T, Jin H, Matese JC, Nitzberg M, Wymore F, Zachariah ZK, Brown PO, Sherlock G (2005) The Stanford Microarray Database accommodates additional microarray platforms and data formats. Nucleic Acids Res Database Issue 33:D580–D582
18 Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R (2005) NCBI GEO: mining millions of expression profiles–database and tools. Nucleic Acids Res Database Issue 33:D562–D566
19 Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B, BirchmeierW, Lehrach H, Wanker EE (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122:957–968
20 Lehner B, Fraser AG (2004) A first-draft human protein-interaction map. Genome Biol 5:R63
21 Turner FS, Clutterbuck DR, Semple CA (2003) POCUS: mining genomic sequence annotation to predict disease genes. Genome Biol 4:R75
22 Brunner HG, van Driel MA (2004) From syndrome families to functional genomics. Nat Rev Genet 5:545–551
23 Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, Clarke L, Coates G, *et al* (2004) An overview of Ensembl. Genome Res 14: 925–928
24 Beaumont MA, Rannala B (2004) The Bayesian revolution in genetics. Nat Rev Genet 5:251–261
25 Egmont-Petersen M, Feelders A, Baesens B (2005) Confidence intervals for probabilistic network classifiers. Comput Stat Data Anal 49:998–1019
26 Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 302:449–453
27 Lee I, Date SV, Adai AT, Marcotte EM (2004) A probabilistic functional network of yeast genes. Science 306:1555–1558
28 Xia Y, Yu H, Jansen R, Seringhaus M, Baxter S, Greenbaum D, Zhao H, Gerstein M (2004) Analyzing cellular biochemistry in terms of molecular networks. Annu Rev Biochem 73:1051–1087 29. Friedman N, Geiger D, Goldszmidt M (1997) Bayesian network classifiers. Mach Learn 29:131–163
30 Jansen R, Gerstein M (2004) Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. Curr Opin Microbiol 7:535–545
31 Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P (2004) Coexpression analysis of human genes across many microarray data sets. Genome Res 14:1085–1094
32 Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: a fast search method for large DNA databases. Genome Res 11:1725–1729
33 Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A (2005) Reverse engineering of regulatory networks in human B cells. Nat Genet 37:382–390
34 Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423, 623–356
35 Floyd RW (1962) Algorithm 97: shortest path. Commun ACM 5: 345
36 Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res Database Issue 33: D514–D517

37 Jain A, Zongker D (1997) Feature selection: evaluation, application and small sample performance. IEEE Trans Pattern Anal 19:153–158

38 Waller WG, Jain AK (1978) Monotonicity of performance of Bayesian classifiers. IEEE Trans Inform Theory 24:392–394

39 Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB (2004) A gene atlas of the mouse and human protein- encoding transcriptomes. Proc Natl Acad Sci USA 101:6062– 6067

40 Stegmaier K, Ross KN, Colavito SA, O'Malley S, Stockwell BR, Golub TR (2004) Gene expression-based high-throughput screening (GE-HTS) and application to leukemia differentiation. Nat Genet 36:257–263

41 Rieger KE, Hong WJ, Tusher VG, Tang J, Tibshirani R, Chu G (2004) Toxicity from radiation therapy associated with abnormal transcriptional responses to DNA damage. Proc Natl Acad Sci USA 101:6635–6640

42 Rieger KE, Chu G (2004) Portrait of transcriptional responses to ultraviolet and ionizing radiation in human cells. Nucleic Acids Res 32:4786–4803

43 Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5:101–113

44 Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, *et al* (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415:180–183

45 Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, *et al* (2003) A protein interaction map of Drosophila melanogaster. Science 302:1727–1736

46 Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, *et al* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415:141–147

47 Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, *et al* (2005) Towards a proteome-scale map of the human protein-protein interaction network. Nature 437: 1173–1178

48 Vidalain PO, Boxem M, Ge H, Li S, Vidal M (2004) Increasing specificity in high-throughput yeast two-hybrid experiments. Methods 32:363–370

49 Egmont-Petersen M, de Jonge W, Siebes A (2004) Discovery of regulatory connections in microarray data. In: Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp 149–160

50 Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, *et al* (2005) InterPro, progress and status in 2005. Nucleic Acids Res Database Issue 33:D201–D205

51 Yandell MD, MajorosWH (2002) Genomics and natural language processing. Nat Rev Genet 3:601–610

52 Malik R, Siebes A (2005) CONAN: an integrative system for biomedical literature mining. Lect Notes Artif Intell 3808:248– 259

53 Bader JS, Chaudhuri A, Rothberg JM, Chant J (2004) Gaining confidence in high-throughput protein interaction networks. Nat Biotechnol 22:78–85

54 Franke L, van Bakel H, Diosdado B, van Belzen M, Wapenaar M, Wijmenga C (2004) TEAM: a tool for the integration of expression, and linkage and association maps. Eur J Hum Genet 12: 633–638

55 Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG (2004) Genetic analysis of genome-wide variation in human gene expression. Nature 430:743–747

56 Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, Friend SH (2003) Genetics of gene expression surveyed in maize, mouse and man. Nature 422:297–302

57 Dudbridge F, Koeleman BP (2004) Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. AmJHum Genet 75:424–435

Figure S1

Difference in likelihood ratio between genes that were represented on the microarrays and genes that were not

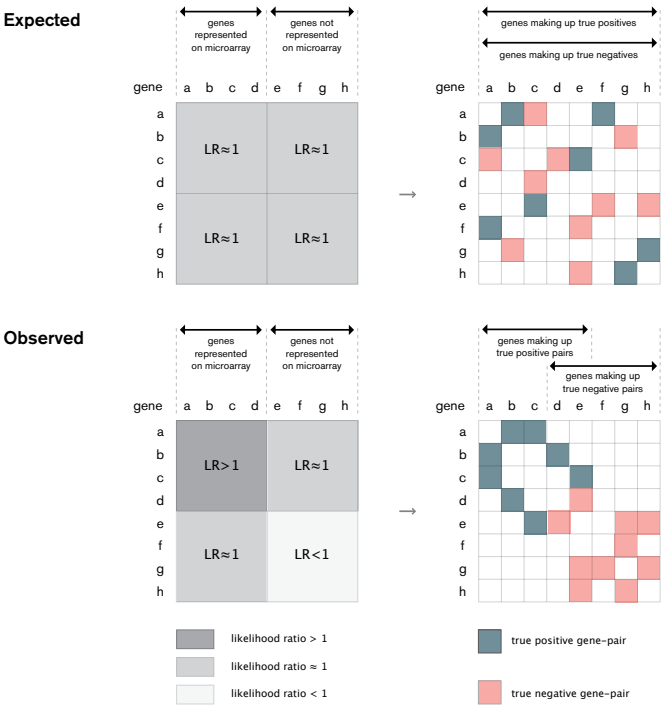
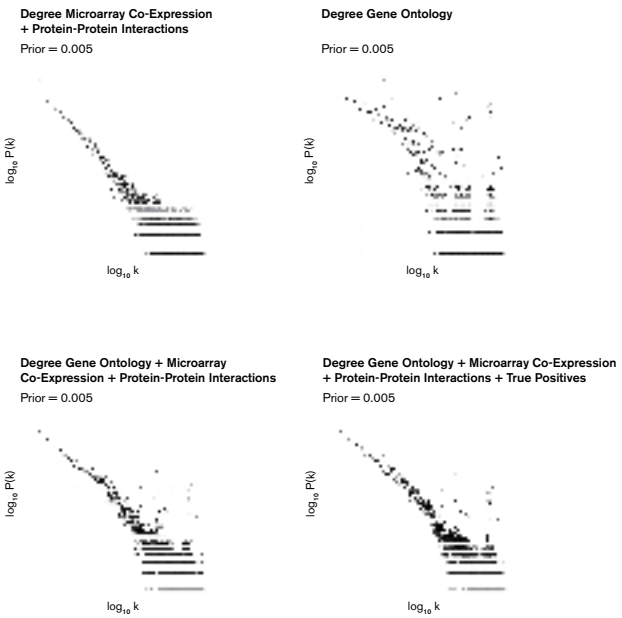


Figure S2

The MA + Y2H network has a topology that most closely follows a scale-free, power-law distribution, compared to the other three networks.



Appendix

Correction for bias in gold standard

Once the likelihood ratios of the microarray co-expression datasets was determined, it became evident that the likelihood ratio of the gene-pairs composed of genes that were present on the microarrays was higher than 1, whereas gene-pairs composed of genes not represented on the arrays was lower than 1. We had not expected there to be any difference in likelihood ratio for genes represented on the array or not, but in fact the observed difference was pronounced (Figure S1).

It turned out that only a small subset of different genes, (7,197 of the 55,606 genes), made up all the 55,606 true positive, gold standard gene-pairs. After we had determined the subset of genes making up the initial, true negative, gold standard gene-pairs and compared this to the 7,197 genes making up the true positive gene-pairs, we found the overlap for individual genes was smaller than expected. A large number of the genes in the true negative gene-pairs were never part of a true positive gene-pair. In addition, the genes in the true negative pairs were generally less well annotated. When we subsequently determine the likelihood ratio of the bins in a dataset, where one bin contains many gene-pairs composed from only a limited number of genes and the other bins contain hardly any gene-pairs formed with one of those genes, the inclusion or exclusion of any of those genes will bias the likelihood ratio, as observed in the microarray co-expression datasets.

To overcome this bias, we tried to come up with a gold standard in which every gene forms both true positive gene-pairs (together with other genes) and also true negative gene-pairs (in conjunction with other, different genes). Only 5,105 genes met our criterion of forming a true positive gene-pair at least three times and these were then allowed to form gene-pairs within the true negative, gold standard reference. This restriction confined the true negative, gold standard reference list to 801,108 gene-pairs, from which 500 gene-pairs could be removed since they were already known to be true positive gene-pairs.

Bayesian integration and graph-theoretical distance measure calculation

To generate the four networks, the various datasets were combined in a Bayesian manner. First, two measures, derived from the GO Biological Process dataset, were combined in a naïve way. This method was also applied to the two GO Molecular Function datasets. Subsequently the overall GO Biological Process dataset and the overall GO Molecular Function dataset were combined in a fully-connected way: for each gene-pair we determined the combination of the two GO bins to which they belonged. Once all the gene-pairs had been assessed, the bin combinations contained a large number of gene-pairs, which permitted determination of the likelihood ratio. Once the overall GO dataset had been generated in a fully connected way, all remaining datasets were combined in a naïve way, as shown in figure 2.

To determine the overall likelihood ratio when combining n datasets f defined as:

$$LR(f_1 \ldots f_n) = \frac{P(f_1 \ldots f_n \mid pos)}{P(f_1 \ldots f_n \mid neg)}$$

we assumed conditional independence between the datasets and used the simplified naïve Bayes formula to compute the likelihood ratio:

$$LR(f_1 \ldots f_n) = \prod_{i=1}^{n} LR(f_i)$$

In order to calculate the eventual microarray co-expression + protein-protein interaction likelihood ratios, the previously determined likelihood ratios from the microarray co-expression dataset and the protein-protein interaction dataset were multiplied. The likelihood ratios of the overall GO dataset and the overall microarray + protein-protein interaction dataset were multiplied to generate the combined GO, microarray co-expression and protein-protein interaction network.

Usually, after the likelihood ratios have been determined, we want to calculate a posterior probability of interaction by multiplying the likelihood ratio with the prior probability of interaction, defined as:

$$O_{posterior} = LR(f_1 \ldots f_n) \cdot O_{prior}$$

However, since it is difficult to estimate the number of existing human gene-gene interactions, no specific prior probability for interaction was assumed. To facilitate successful learning of the classifier, a uniform (ignorant) prior was used for both the training and test set, which allowed us to adjust the computed posterior probability to take into account any plausible prior probability of interaction.

In addition, it was decided not to discretize the gene-pairs into interacting or non-interacting pairs, but to use a continuous graph-theoretical distance measure which could be employed in the graph network, while taking into account that evidence for interaction could vary. First, all gene-pairs were ranked based on the computed likelihood ratio. Subsequently this ranking was used to define a distance measure which ranged from one (highly likely gene-gene interaction) to 255 (highly unlikely gene-gene interaction) and followed the following cumulative distribution function (CDF):

$$CDF(distance) = \frac{49}{50} \cdot \left(\frac{20334 \cdot (20334 - 1)}{2} \right)^{\frac{distance}{255}} + \frac{distance}{50 \cdot 255}$$

Degree distribution of the four networks

Degree distributions were determined to assess whether the reconstructed networks followed a scale-free, power-law distribution (Figure S2). The number of interacting genes per gene k was determined, assuming a conservative prior of 0.005. The proportion p of genes having *k* interactions was plotted against *k*.

3 Detection, imputation and association analysis of small deletions and null-alleles on oligonucleotide arrays

American Journal of Human Genetics, in press

Lude H. Franke [1,3], Carolien G.F. de Kovel [1], Yurii S. Aulchenko [2], Gosia Trynka [3], Alexandra Zhernakova [1], Karen A. Hunt [4], Hylke M. Blauw [5], Leonard H. van den Berg [5], Roel A. Ophoff [1,6], Panagiotis Deloukas [7], David A. van Heel [4], Cisca Wijmenga [1,3]

1 Complex Genetics Section, DBG-Department of Medical Genetics, University Medical Centre Utrecht, 3584 CG Utrecht, the Netherlands
2 Department of Epidemiology & Biostatistics, Erasmus MC Rotterdam, 3000 CA Rotterdam, the Netherlands
3 Genetics Department, University Medical Centre Groningen and University of Groningen, 9700 RB Groningen, the Netherlands
4 Institute of Cell and Molecular Science, Barts and The London School of Medicine and Dentistry, London, E1 2AT, UK
5 Department of Neurology, Rudolf Magnus Institute of Neuroscience, University Medical Center Utrecht, 3584 CX Utrecht, the Netherlands
6 Center for Neurobehavioral Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA 90095, USA
7 Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK

Summary

Copy number variation (CNV) is a major contributor to human genetic variation. Recently, CNV associations with human disease have been reported. Many genome-wide association (GWA) studies in complex diseases have been performed using sets of biallelic single nucleotide polymorphisms (SNPs), but the available CNV methods are still limited. We present a new method (TriTyper) that can infer genotypes in case-control datasets for deletion CNVs, or SNPs with an extra, untyped allele at a high-resolution single SNP level. By accounting for linkage disequilibrium (LD), as well as intensity data, calling accuracy is improved. Analysis of 3,102 unrelated Caucasian individuals, genotyped using Illumina Infinium BeadChips, resulted in the identification of 1,880 SNPs with a common untyped allele that are in strong LD with neighboring biallelic SNPs. Simulations indicate our method has superior power to detect associations compared to biallelic SNPs that are in LD with these SNPs, yet without increasing Type I errors, as shown in a GWA analysis in celiac disease. Genotypes for 1,204 triallelic SNPs could be fully imputed, using only biallelic genotype calls, permitting association analysis of these SNPs in many published datasets. We estimate that 682 of the 1,655 unique loci reflect deletions; this is on average 99 deletions per individual, four times more than detected by other methods. Whilst the identified loci are strongly enriched for known deletions, 61% have not been reported before. Genes overlapping with these loci more often have paralogs ($P = 0.006$) and biologically interact with fewer genes than expected ($P = 0.004$).

Introduction

It has become apparent that copy number variation (CNV) accounts for a considerable amount of genetic variation[1-5] and has been implicated as a causal mechanism for several disorders[6-8]. Specialized comparative genomic hybridization (CGH) arrays that contain large-insert clones that hybridize to complementary DNA[1; 5; 9; 10] have provided much insight into the properties of CNVs. These studies have shown that individuals usually carry many small deletion- and duplication CNVs that can be found with high population frequencies.

Recently, much effort has been devoted to detecting CNVs using single nucleotide polymorphism (SNP) genotype data in both familial and unrelated samples[2; 4; 11-19]. An important resource so far has been the HapMap project[20], in which over three million SNPs have been typed for 270 samples. In addition, growing resources of genotype data from oligonucleotide arrays that usually assay at least 300,000 SNPs have been generated for genome-wide association (GWA) studies. Although there are technical challenges to detecting CNVs using these arrays[21], various methods have been developed. Some have been designed to work on single samples[13; 14; 17-19; 22], using similar principles as used for array CGH, while others take multiple samples jointly into consideration[2; 4; 15; 22]. The single sample methods typically require that multiple, consecutive (usually at least three) SNPs show deviations in the allele intensity signals. When multiple samples are analyzed together, genotype calls, based on biallelic SNP assumptions, can provide circumstantial evidence that CNVs span these SNPs. SNPs that map within common CNVs are expected to show deviations from Hardy-Weinberg equilibrium (HWE) and an increased number of missing genotype calls. If family data is present, a control for Mendelian segregation is routinely performed. Usually this is done to determine genotyping accuracy, but if for a given SNP segregation inconsistencies are observed, these can also be caused by violations of the assumption that the SNP is biallelic: duplications, deletions, or the presence of a third allele at the locus that is not labeled by the assay, can all lead to observations of Mendelian inconsistency.

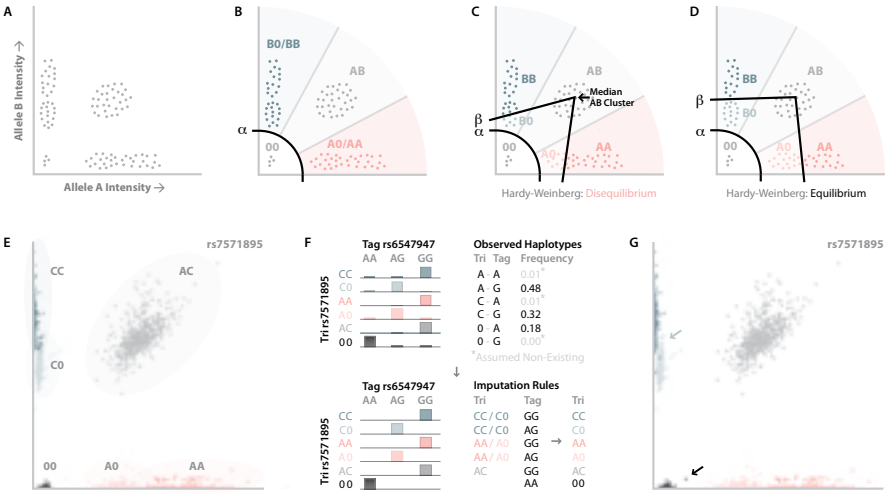
One limitation of the available CNV detection methods is the resolution, as nearly all require that multiple consecutive SNPs show aberrant intensity characteristics[4; 13; 14; 16-19; 22]. One method has a resolution as high as a single SNP[15], but can only be applied to families.

Here we describe a new genotype-calling method (*'TriTyper'*) that can reliably detect deletions in unrelated samples that span only one SNP. Our algorithm detects SNPs with an extra, untyped allele (including deletion CNVs encompassing these SNPs) using raw intensity data from Illumina® Infinium HumapHap300 and HumanHap550 BeadChip arrays[23]. Using *TriTyper* we identified 1,880 SNPs with a common extra allele (frequency > 0.5%) in a collection of 3,102 DNA samples from individuals of Northwest European origin. Our method can accurately assign genotypes by utilizing local linkage disequilibrium (LD) with nearby SNPs[1; 24; 25]. We show that our procedure results in correct genotype assignments through a Mendelian segregation analysis in Caucasian HapMap trios, where many segregation inconsistencies, observed under biallelic-calling assumptions, are resolved when triallelic genotypes have been assigned. Of the 1,880 triallelic SNPs, 1,204 can be fully imputed from surrounding SNPs without the need to use raw intensity data. This is helpful when analyzing triallelic SNPs in publicly available and other datasets for which only genotype calls have been made available. We show how these triallelic genotypes can be used for association studies and that our test statistic shows no inflation in significant signals as exemplified in an analysis of celiac disease. Yet, like other imputation methods[26; 27], our method has superior power to detect true positive associations, when contrasted to an association analysis of nearby biallelic SNPs, used for imputing the triallelic SNPs. The identified triallelic loci are strongly enriched for known deletions, but the majority of identified deletions have not yet been described. We support previous findings that genes, mapping within these deletions, more often have paralogs, but we

Figure 1

Genotyping methodology for SNPs with a third, untyped allele

The graphs show the intensities of the A-labeled probe (x-axis) and B-labeled probe (y-axis) of both a theoretical SNP with an third, untyped allele (top figures) and a real SNP (rs7571895, bottom figures). **a)** Six genotypes for a triallelic SNP exist. The A0 and AA, and B0 and BB, genotype clusters usually overlap somewhat. **b)** Initially 00 genotypes are assigned to samples that have an intensity lower than threshold α. The remaining samples are designated an initial A0/AA, AB or B0/BB genotype using an existing calling algorithm. **c)** Parameter β is then used to discriminate between A0 and AA and between B0 and BB genotypes (see text). This allows for determining whether Hardy-Weinberg equilibrium is observed. **d)** Parameters α and β are then optimized (using a maximum likelihood estimation procedure) until the SNP does adhere to Hardy-Weinberg equilibrium conditions. **e)** Triallelic genotype assignments, based on the MLE procedure for SNP rs7571895, are shown. **f)** Subsequent analysis of neighboring SNPs results in the identification of biallelic SNP rs654797, which is in strong LD with the null-allele of rs7571895. Although LD does not seem to be perfect ($r^2<1$) we assume this is likely due to imperfections in the initial genotype assignments, and that some of the haplotypes (indicated with an asterisk) are not actually present. This allows for identifying a set of triallelic genotype imputation rules that are applied to the data and result in **g)** improved genotype assignments for rs7571895, as is clearly visible when distinguishing C0 from CC samples (green arrow) and 00 from A0 samples (black arrow).



also found that the genes usually tend to interact biologically with fewer genes than expected. By using *TriTyper*, more genetic information can be captured, triallelic SNP genotypes can be imputed, and interesting phenomena, including small deletion CNVs, can be detected in numerous case-control cohorts that have already been typed on oligonucleotide platforms.

Materials and Methods

Triallelic genotype-calling algorithm

Oligonucleotide assays, available for high-throughput SNP genotyping, usually measure the intensities of two fluorescent labels that are attached to two known alleles, A and B. Throughout this paper these are plotted on the x-axis (intensity$_a$) and y-axis (intensity$_b$), respectively. When an extra, untyped allele (a 'null' or 0 allele) is present, up to six clusters (representing AA, AB, BB, A0, B0 and 00 genotypes) in the raw intensity plot will become visible (fig. 1A). Usually these A0 and B0 clusters partly overlap with the AA and BB clusters, respectively, while the 00 cluster has a very low Euclidian intensity. We refer to this as a "triallelic" pattern or "triallelic" SNP. If the presence of this null-allele is not recognized, standard calling algorithms will typically call A0 and B0 genotypes as AA and BB respectively, and 00 genotypes as 'failed'. Under biallelic assumptions, deviations from HWE are then likely to become apparent.

We used these deviations under biallelic assumptions as the basis for our triallelic genotype-calling algorithm (*TriTyper*). *TriTyper* extends a biallelic genotype-calling algorithm we have recently developed[28] and models triallelic genotypes by using a maximum likelihood estimation (MLE) procedure that optimizes HWE under triallelic assumptions[29] (fig. 1A-D, for details see appendix A). Another key aspect of our method is that it uses the presence of local LD between this null-allele and nearby biallelic SNPs[1; 24; 25] to gain evidence that the extra allele has been correctly identified. Once this has been established, it takes advantage of these biallelic SNPs to improve the triallelic genotype assignments by using a fairly straightforward imputation method (fig. 1E-G, for details see appendix A) that borrows some ideas from methods that impute genotypes for biallelic SNPs[26; 27]. This imputation methodology often allows for accurately discriminating between A0 and AA and between B0 and BB samples, which is particularly helpful as these clusters usually overlap somewhat (fig. 1G, green arrow).

Datasets for triallelic SNP discovery

Initial analyses were performed on a cohort that comprised 1,422 unrelated control individuals[28] from the 1958 British Birth Cohort that passed quality control (QC) and had been typed on the Illumina Infinium II Human Hap550 BeadChip platform for 571,738 SNPs. To also detect triallelic SNPs with lower null-allele frequencies, we added three more cohorts. These included 778 unrelated UK celiac disease cases[28], 450 unrelated Dutch controls[30] and 472 unrelated Dutch amyotrophic lateral sclerosis cases[30] that all passed QC and had been typed on the Illumina Infinium II Human Hap300 BeadChip platform for 317,503 SNPs. In this combined analysis 313,505 SNPs could be analyzed, as they were present on both the Hap300 and Hap550 platforms. 20 samples (0.6%) showed aberrant intensity signals for many of the triallelic SNPs and were removed from the analyses.

Association analysis

An analysis for marginal association effects on the biallelic SNPs used for imputation of the triallelic SNPs was performed as follows: analyses were confined to SNPs for which the null-allele was not in complete LD with a biallelic SNP, as for these SNPs Fisher's Exact test for association would be identical to the association analysis of the triallelic null-allele. Only triallelic SNPs, where one biallelic SNP could help to discriminate between A0 and AA genotypes and another biallelic SNP could help to discriminate between B0 and BB genotypes, were included in the analysis.

To assess the marginal effect on the SNPs used for imputing the triallelic SNPs, three different scenarios of triallelic SNP association (Fisher's Exact test for the triallelic null-allele of 10^{-4}, 10^{-6} and 10^{-8}) were simulated. For each triallelic SNP an equal number of controls and cases were chosen, but case and control labels were assigned in such a

way that association for the triallelic SNP yielded a Fisher's exact P value for the null-allele that approximated the P value of the scenario under investigation. This allowed for determining the marginal association effect on the two biallelic SNPs used for imputing each triallelic SNP. To gain accurate estimates this was repeated 100 times. Subsequently, for each triallelic SNP, the average marginal effect on the biallelic SNP that was associated most significantly was recorded. Once this was performed for all the triallelic SNPs, the median marginal effect could be determined for each scenario.

The triallelic SNP null-allele association analysis was performed on a celiac disease GWA dataset[28] and was confined to those triallelic SNPs for which imputation could help to discriminate between both the A0 and AA samples and between the B0 and BB samples. We did this as different arrays had been used to genotype cases and controls. Although these arrays for most SNPs show highly comparable intensity characteristics, for some SNPs subtle differences are present. When nearby biallelic SNPs can only help to discriminate between A0 and AA or between B0 and BB, spurious associations are to be expected due to the way our calling algorithm initially discriminates between A0 and AA, and B0 and BB genotypes. Because of the normally low frequency of the null-allele, a Fisher's Exact test was performed for testing the association significance. Type I errors were ascertained by a quantile-quantile (Q-Q) plot, generated by plotting the observed ordered null-allele associations against the ordered expected associations. Then we fitted a line to the lower 90% of the distribution, of which the slope ($\lambda_{inflation}$) denotes either the inflation or deflation of the test statistic.

Segregation analysis

A segregation analysis was performed on 16 CEU trios for which biallelic genotype data had been generated on the Illumina Infinium II Human Hap650 platform (containing 660,918 SNPs). We chose this dataset since no genotypes for many of the identified triallelic SNPs were available in the Phase II release from HapMap; this was due to the fact that SNPs showing segregation inconsistencies in multiple trios were not included in this release.

Triallelic SNPs were included for analysis if genotypes could be imputed based on the biallelic calls, thus without directly relying upon the raw intensity data, requiring that genotype calls for these SNPs and the biallelic SNPs used for imputation were available. Imputation allowed us to inspect visually whether the raw intensity data patterns corresponded well to the imputed genotype assignments. Subsequently, we used these imputed triallelic genotypes to assess how many of the Mendelian segregation inconsistencies observed under biallelic assumptions could be resolved. We took a conservative approach, because we did not score segregation inconsistencies in the analysis of the biallelic genotype calls in trios in which a genotype had not been called for either the mother or the father.

Identity of untyped alleles

Various sources can result in the detected null-alleles within the identified triallelic SNPs. Deletion CNVs that span these SNPs will result in these triallelic intensity characteristics, while a previously unknown, third nucleotide at the physical position of the SNP gives the same results. Alternatively, it is possible that within the immediately adjacent locus which is complementary to the 50 bp primer of the SNP (used in the Illumina Infinium chemistry), there is a secondary polymorphism that affects the hybridization efficacy of the primer and that will consequently result in the same triallelic pattern[31].

To discriminate between these three possible explanations, we investigated whether there was any evidence that these SNPs reside within deletion CNVs. If a deletion CNV is large enough to span multiple assayed SNPs, these SNPs should all show a triallelic intensity characteristic. It is likely they will all be identified by our calling method, but some might be missed (Type II error). To overcome this, for each triallelic SNP we assessed whether its neighboring SNPs showed characteristics suggesting the presence of a triallelic pattern. It is expected that if this is the case, a neighboring SNP

(like the triallelic SNP) will show Euclidian intensities for the triallelic A0 and B0 samples that are significantly lower than the intensities of the samples with a triallelic AA, AB or BB genotype.

We first corrected for differences in probe intensity characteristics within these neighboring SNPs through ranking the Euclidian intensities of the samples that had an AA genotype for the neighboring SNP and through ranking the Euclidian intensities of the samples that had a BB genotype for the neighboring SNP. We linearly scaled these two rankings to [0, 1] and assigned a value of 0.5 to samples that were heterozygous for the neighboring SNP. We then compared the ranked intensities of the samples that had been assigned triallelic 00, A0 or B0 genotypes with the ranked intensities of samples with triallelic AA or BB genotypes, and required that ranked intensities of the 00, A0 and B0 samples were significantly lower (one-sided Wilcoxon-Mann-Whitney test P value $< 10^{-5}$). We then called genotypes under biallelic assumptions for the neighboring SNP. We also required that loss-of-heterozygosity (LOH) was observed (Fisher's Exact test P value < 0.01) in the samples that had been assigned 00, A0 or B0 genotypes for the triallelic SNP. However, we only tested for this if the minor allele frequency of the neighboring SNP was high enough, such that in a theoretical situation where no AB samples were present, the LOH Fisher's Exact test P value would be below 0.001.

We first performed this analysis for the immediately adjacent SNPs and then moved further to the left and right, continuing as long as the above conditions applied. As the A0 and AA clusters and B0 and BB clusters usually overlap somewhat, we reasoned that if a deletion spans several SNPs, a better separation between A0 and AA samples and between B0 and BB samples would be obtained if we averaged the ranked intensities of these SNPs per sample. We applied this as an extra criterion for determining how far a deletion is likely to extend. Apart from the above criteria, we also required that, when we included more neighboring SNPs to the left and right of the triallelic SNP, the averaged ranked intensity differences between the samples with an A0 or B0 genotype and the samples with an AA and BB genotype should consistently become more significant.

These criteria meant we could determine the locus size for each fitted triallelic SNP. Immediately overlapping and adjacent loci were concatenated, resulting in loci that ranged in size between one SNP and loci that contained multiple fitted SNPs and/or neighboring SNPs that showed aberrant intensity characteristics and LOH.

To identify SNPs for which the observed triallelic intensity characteristic was due to a polymorphism in the primer region, we derived the physical genomic positions where the 50 bp primers anneal and determined whether more polymorphisms had been described within these loci in dbSNP (build 127). All analyses were performed on the NCBI build 36 genome assembly.

All the triallelic loci identified were categorized into loci that contained multiple consecutive triallelic SNPs, loci that contained one SNP for which no polymorphisms within the primer were known, and loci that contained one single triallelic SNP and for which a primer polymorphism was known.

Resequencing

We selected 23 triallelic SNPs for resequencing. Two were selected to corroborate our prediction that the null-allele for these was caused by primer polymorphisms. An additional 21 triallelic SNPs were selected to get an estimate of what proportion of the identified null-alleles reflects primer polymorphisms and what proportion reflects deletions. To assess the quality of the genotype predictions, we selected triallelic SNPs with different inferred genotype qualities. We selected samples for all six genotypes when possible. Primers were designed such that we PCR amplified approximately 500 base pairs around the triallelic SNPs. On average nine samples were sequenced per SNP. Sequencing was performed according to standard protocols on an ABI 3730 (Applied Biosystems) sequencer.

Genomic properties of triallelic loci

Ensembl[32] version 41.36c was used for annotation purposes and mapping of gene identifiers to Ensembl gene names. The size of each identified locus was defined by taking the physical distance between the two immediate biallelic SNPs that enclosed it. The significance of under- or overrepresentations for each of the various genomic properties was empirically determined by permuting all loci across the genome 1,000 times, through defining the loci randomly around SNPs that were present on the Illumina Hap550 chip and ensuring that the size of these permuted loci was equal to the real distribution. Known deletion CNVs were derived from the Database of Genomic Variants[3] (March 2007 release, NCBI build 36 mapping). Enrichment of the loci for these deletions was assessed by determining how many loci overlapped with known deletion CNVs and by fitting an extreme value distribution (EVD) on the permuted loci using the EVD add-on package[33] to R (R Development Core Team 2003, version 2.4.1). The Online Mendelian Inheritance in Man[34] morbid map (downloaded on 6 December 2006) was used for the enrichment analysis of disease genes that overlapped with our loci. Enrichment analysis of genes with known paralogs was determined empirically by deriving all known paralogs from Ensembl and assessing whether the number of genes that overlapped with the identified loci with known paralogs was higher than within the permutations. Known biological interactions were derived from KEGG[35], BioGrid[36], Reactome[37], BIND[38], HPRD[39] and IntAct[40] (all downloaded on 17 April 2007). Interaction depletion analysis for the genes, overlapping with the identified loci, was determined by contrasting the distribution of the number of interactions ('degree') for each of these genes against the distribution of the degree of the genes that were present within the 1,000 permutations, using a Wilcoxon-Mann-Whitney test.

Results

Identification of 1,880 triallelic SNPs

TriTyper initially determines which SNPs show deviation from HWE under biallelic assumptions, which provides evidence that an extra, untyped allele might be present for these SNPs (see figure 1A and details in appendix A). For these SNPs we tried to fit 'triallelic' genotypes (fig. 1A, see details in appendix A). Initially we used parameter α to identify a putative set of samples with 00 genotypes, and assigned preliminary A0/AA, AB and B0/BB genotypes to the remaining samples (fig. 1B). We used parameter β to distinguish both between A0 and AA samples and between B0 and BB samples (fig. 1C). By adjusting α and β, and using a maximum likelihood estimation procedure, we could then find a triallelic genotype assignment where HWE was observed (fig. 1D). We then looked for circumstantial evidence that this untyped allele had been correctly identified (fig. 1E) by searching nearby biallelic SNPs that are in near perfect LD with this null-allele (fig. 1F). As some of the initially assigned genotypes might be incorrect, we can use this LD to improve upon the triallelic genotyping through imputation (fig. 1G, green and black arrows) (see details in appendix A).

By applying this algorithm to 1,417 unrelated UK controls, genotyped for 571,738 SNPs (Illumina Human Hap550 array), we identified 1,535 triallelic SNPs (median null-allele frequency = 8.6%). To be able to detect triallelic SNPs with a lower null-allele frequency we increased the sample size to 3,102, by adding 768 unrelated UK celiac patients, 445 unrelated Dutch controls, and 472 unrelated Dutch amyotrophic lateral sclerosis patients. As these samples had been typed on the Illumina Human Hap300 array, this analysis was restricted to the 313,505 SNPs that were present on both array types. We identified 958 triallelic SNPs, of which 345 (median null-allele frequency = 4.7%) had not been identified in the smaller cohort. Cluster plots of all 1,880 triallelic SNPs are available on the *TriTyper* website.

The presence of LD between these null-alleles and nearby biallelic SNPs provides strong evidence that an untyped allele has been correctly identified for these triallelic SNPs. In addition, once the presence of this LD had been established we utilized it to partly impute the triallelic genotypes. For 1,204 (64%) of the 1,880 triallelic SNPs, imputation is capable of discriminating both

between A0 and AA and between B0 and BB samples. In these cases biallelic genotype calls suffice to infer these 'fully imputable' triallelic genotypes. This allows for performing association analysis of triallelic SNPs in GWA studies for which only biallelic genotype calls have been made publicly available[41; 42], or when different genotyping assays have been used.

To assess how well imputation functions when only biallelic genotype calls and no raw intensity data were available, we performed a Mendelian segregation analysis on genotype data from 16 CEU trios. For these samples biallelic genotype calls were available for 1,153 (96%) of the 1,204 fully imputable triallelic SNPs (see "Material and Methods" section). 431 (37%) SNPs showed segregation inconsistencies under biallelic assumptions. When imputing triallelic genotypes, this decreased to 319 (28%). This indicates that some segregation inconsistencies can indeed be resolved. We reasoned that if the LD was high between the null-allele and the biallelic SNPs used for imputation, the genotypes should mostly be correct and would resolve most of the observed segregation inconsistencies. To assess this we confined the analysis to those triallelic SNPs in our cohort for which the observed concordance between the preliminary triallelic genotypes determined and the subsequently imputed triallelic genotypes was at least 90%. Of these 596 triallelic SNPs, 257 (43%) showed Mendelian segregation inconsistencies when called under biallelic assumptions, compared to 60 (10%) when using the imputed triallelic genotypes (individual segregation plots are available at the *TriTyper* website). This implies that for the great majority of the identified SNPs, an extra allele has indeed been typed but that most of these triallelic genotypes can be correctly imputed when the LD is sufficiently high. Additionally, the concordance between the preliminary assigned triallelic genotype and eventually imputed genotypes serves as a quality statistic measure of the triallelic genotype calling.

Association analysis

As most GWA studies aim to identify new susceptibility loci for diseases, it is essential that accurate association analysis can also be performed on the triallelic SNPs identified. We first investigated whether such an analysis has higher statistical power than an analysis of biallelic SNPs that are in LD with these triallelic SNPs, as we expected some marginal effect on these nearby biallelic SNPs to be observed as well. To assess the strength of this marginal effect, we simulated null-allele associations for 600 triallelic SNPs under three association scenarios (Association $P = 10^{-4}$, $P = 10^{-6}$ and $P = 10^{-8}$, see Material and methods section). For each scenario, case and control labels for each triallelic SNP were assigned in such a way that the association P value for the null-allele of this SNP approximated the P value of the scenario under investigation. Then the association strength of the SNPs used for imputation purposes could be determined (fig. 2A). The median marginal effect was $3*10^{-3}$, $3*10^{-4}$, $2*10^{-5}$ for the three scenarios, respectively, indicating that marginal effects on the SNPs used for imputation are usually present, but much weaker than for the imputed triallelic SNP. It can thus be concluded that the statistical power to detect associations for the null-alleles of these triallelic SNPs is considerably higher than an analysis of the biallelic SNPs that are in LD with them.

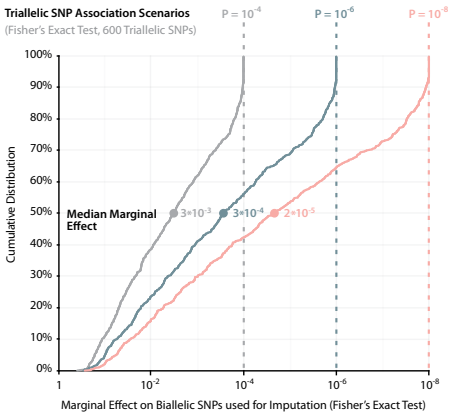
We performed a celiac disease association analysis on the triallelic SNPs identified in the dataset[28] that comprised 1,417 UK controls and 768 celiac disease cases. Celiac disease is a common (1% prevalence), inflammatory condition of the small intestine induced by intake of gluten in wheat, rye and barley. Most of the heritability is explained by the human leukocyte antigen (HLA) component[43], as the majority of individuals with celiac disease possess HLA-DQ2 (and the remainder mostly have HLA-DQ8)[44]. Recently, we identified additional susceptibility loci in a GWA study[28; 45; 46], in which we performed an association analysis on 585 fully imputable triallelic SNPs (see Material and methods section). The results (fig. 2B) indicate that an association analysis on these triallelic SNPs does not lead to inflated test statistics, as $\lambda_{inflation} = 0.96$ when calculated on the lower 90% of the distribution ($\lambda_{inflation} = 1.08$ when calculated with all test statistics). This suggests that our imputation methodology prevents spuri-

Figure 2

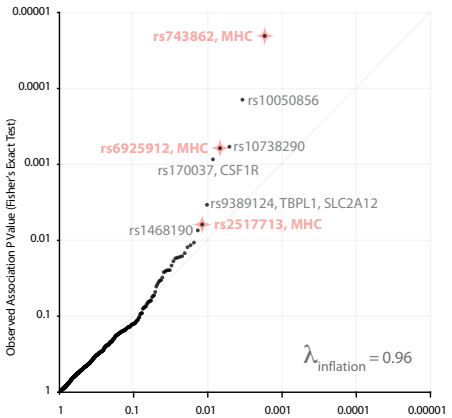
Association analysis using triallelic SNPs and marginal effect on SNPs, used for imputation

a) Marginal association signals of SNPs, used for imputing triallelic SNPs, with disease. Fixed associations for the null-allele of 600 triallelic SNPs were defined in such a way that each of the triallelic SNPs approximated a Fisher's Exact Test P value of 10^{-4}, 10^{-6} or 10^{-8}. We then assessed whether a marginal association signal was present within the SNPs that had been used to impute the triallelic genotypes. The median marginal effect and the cumulative distribution of the marginal association P value for each of these SNPs are shown, ranked on significance (see text for details). **b)** Quantile-Quantile plot of observed versus expected P values in a triallelic SNP null-allele association analysis in celiac disease, where cases and controls had been typed on different platforms. Eight triallelic SNPs with a Fisher's Exact Test P value < 0.01 are indicated. The $\lambda_{inflation}$ factor is 0.96, suggesting no inflation of the test statistic. Three SNPs map within the major histocompatibility complex region (indicated in red).

a)



b)



ous associations, which is quite encouraging since the cases and controls had been typed on different arrays (Illumina Human Hap300 versus Illumina Human Hap550). Eight triallelic SNPs showed a Fisher's Exact test P value below 0.01 (table 1). When we expanded the control cohort by adding 445 Dutch controls, all eight SNPs retained a P value < 0.01. Three of these (rs743862, rs6925912 and rs2517713, marked red in figure 2B) map within or very close to the major histocompatibility complex (MHC) that is highly polymorphic, has extended LD, and contains the strongly associated *HLA-DQA1* and *HLA-DQB1* genes. As such, these null-alleles probably reflect nearby polymorphisms (located on a celiac disease associated haplotype) that affect the annealing of the triallelic SNP primers. Based on dbSNP (build 127), this is known to be the case for rs743862 (rs28366194 at +1bp) and rs2517713 (rs9260378 at +3 bp). Although such a secondary 'primer polymorphism' is not known for rs6925912, this cannot be excluded as the MHC is highly polymorphic. For the remaining five triallelic SNPs there is little evidence for their potential involvement in celiac disease, with the notable exception of rs170037. This SNP maps within a known susceptibility locus (CELIAC2 on 5q31-33) that has been identified in independent linkage studies[47-49] and was significantly linked in a meta-analysis of four populations[50]. It maps in an intron of the colony stimulating factor 1 receptor (*CSF1R*) that is involved in monocyte to macrophage differentiation and innate immunity[51]. For *CSF1R* some weak association has also been reported with Crohn's disease[52], another inflammatory gastrointestinal disorder for which molecular mechanisms, comparable to celiac disease, have been implicated[46].

It is relevant to note that if the null-allele itself is not associated with disease, but the A or B alleles are, biallelic assumptions will result in either an over- or underestimation of the effect, depending on whether the effect is dominant or recessive, respectively (see details and figure 5). While these triallelic SNPs are usually excluded from biallelic association analyses, due to observed HWE deviations, it is possible these deviations remain under the threshold used (usually in GWA studies an Exact HWE P value < 0.0001 is used to exclude SNPs from subsequent association analysis[28]). This is likely to be the case if the sample size is small, indicating that when associations are observed for any identified triallelic SNP under biallelic assumptions, one should proceed with caution.

Identity of null-alleles

The detected null-alleles within the 1,880 triallelic SNPs can originate from different sources. These SNPs might map within deletion CNVs, which will result in the observed triallelic intensity characteristics, but the null-allele might also reflect an unknown, third nucleotide at the physical position of the SNP (e.g. an A/C SNP in fact is an A/C/G SNP). Another explanation could be that, within the immediately adjacent locus that is complementary to the 50 bp primer of the SNP, a secondary polymorphism is present that affects the hybridization efficacy of the primer and consequently results in the same triallelic pattern[31]. To gain insight into these classes, non-overlapping loci (see fig. 3 and table 2) were defined by concatenating immediately adjacent triallelic SNPs. 208 of the SNPs that were immediately adjacent to the triallelic SNPs, but which had not been deemed triallelic, were also added because they showed aberrant intensity characteristics and loss of heterozygosity (see the Materials and methods section). This resulted in the identification of 1,655 different loci in total.

145 loci spanned multiple adjacent SNPs, which suggests these loci reflect deletions and this is supported by an analysis of the Database of Genomic Variants. 77 (53%) were already known to be deletions in this database, which is much more than expected (Extreme Value Distribution P value $< 10^{-50}$).

For the remaining 1,510 loci that contained only one SNP, the origin of the extra allele was less obvious: one explanation could be that polymorphisms map within the locus that is complementary to the 50 bp primer of the SNP, affecting the hybridization efficacy of the primer and resulting in this triallelic pattern. These primer polymorphisms were observed in 437 (29%) of these loci

Table 1 Triallelic SNPs with null-allele, associated with celiac disease ($P < 0.01$)

Triallelic SNP	Chr.	Position (bp)	Overlapping genes (nearby genes)	Null-allele frequency UK cases	Null-allele frequency UK controls
rs743862 †	6	32,489,917	(BTNL2, HLA-DRA)	12.5%	8.4%
rs10050856 †	5	23,407,397	(PRDM9)	13.5%	9.6%
rs10738290*	9	12,730,906	(TYRP1, C9orf150)	3.5%	5.8%
rs6925912	6	26,084,906	TRIM38	12.0%	15.8%
rs170037	5	149,420,837	CSF1R	4.9%	7.5%
rs9389124	6	134,355,478	TBPL1,SLC2A12	4.4%	6.5%
rs2517713 † *	6	30,026,078	HLA-A	2.7%	4.4%
rs1468190	16	13,265,389	(ERCC4)	17.3%	20.7%

Italic SNPs mapping within major histocompatibility complex (MHC)

† Known polymorphism present within the primer of SNP (dbSNP, build 127)

* Known deletion locus (Database of Genomic Variants, March 2007 release)

Null-allele frequency Dutch controls	Association P value UK samples (Fisher's Exact Test)	Allele frequency P value of SNPs used for imputation on UK samples (1df χ^2 test)	
6.3%	2.02 x 10-5	rs9501626	0.013
		rs3817963	2.63 x 10^{-10}
10.8%	1.40 x 10-4	rs10038792	0.274
		rs3924616	0.0978
5.8%	5.86 x 10-4	rs970946	0.863
		rs391858	0.002
15.8%	6.10 x 10-4	rs199750	9.09 x 10^{-6}
		rs199741	1.70 x 10^{-4}
6.1%	8.58 x 10-4	rs216148	0.028
6.0%	0.0034	rs6902440	0.017
5.8%	0.0061	rs2860580	1.11 x 10^{-16}
		rs2256902	0.005
21.0%	0.0074	rs10492781	0.004

(table 2), which is considerably higher than expected, as secondary polymorphisms are known within the primer region for 85,045 (16%) of the 550,123 Human Hap550 SNPs with known mapping (Fisher's Exact Test P value $< 10^{-18}$). Interestingly, when assessing how far these primer polymorphisms map away from the triallelic SNP, the two distributions showed a markedly different distribution (see figure 6). Primer polymorphisms were usually much closer to the investigated triallelic SNP compared to the distribution of the other SNPs with known primer polymorphisms (Wilcoxon Mann-Whitney P value $< 10^{-76}$). This implies that primers on the Illumina platform usually tolerate polymorphisms well, as long as these do not map too close (> 10 bp) to the SNP to be typed.

For the 1,073 loci without known primer polymorphisms, we observed a strong enrichment of deletions, known in the Database of Genomic Variants, as 136 (13%) had been reported in this database (Extreme Value Distribution P value $< 10^{-50}$). Earlier estimates show that 50%[31] to 60%[5] of these loci reflect deletions. This suggests we have detected at least 682 small-deletion CNV regions (assuming 50% of the 1,073 loci reflect deletions and adding the 145 multiple SNP loci). With an observed median null-allele frequency of 7.6% for these loci, this suggests we have identified 99 deletions per individual on average. A negative binomial distribution fits the observed allele frequency distribution (fig. 4B) well. An exponential distribution fits the observed triallelic locus size distribution (fig. 4A, median size = 7,290 bp), supporting previous observations that small CNVs strongly outnumber larger ones[4; 53].

Resequencing

We resequenced 23 triallelic SNPs to assess the predicted proportion of deletions among the identified triallelic SNPs (table 3). For two triallelic SNPs (rs13213842 and rs7678151) we confirmed that the observed null-allele was indeed due to a primer polymorphism. For the other 21 triallelic SNPs we observed that the null-allele reflects a primer polymorphism in ten SNPs. Small deletions were identified in two SNPs (rs7822381 and rs2486674). For the other

Figure 3

Overview of 1,655 triallelic loci identified on autosomes and chromosome X

Immediately to the left of each chromosome are depicted all the SNPs present on the Illumina Human Hap550 platform. CNVs known in the Database of Genomic Variants are shown to the right of each chromosome. Next to this the triallelic loci are shown where the length of each bar denotes the null-allele frequency. Metallic indicates a single triallelic SNP locus, red a locus in which multiple adjacent triallelic SNPs have been identified, and grey indicates a single triallelic SNP locus for which polymorphisms are known within the region complementary to the primer of the triallelic SNP (dbSNP build 127, March 2007 release).

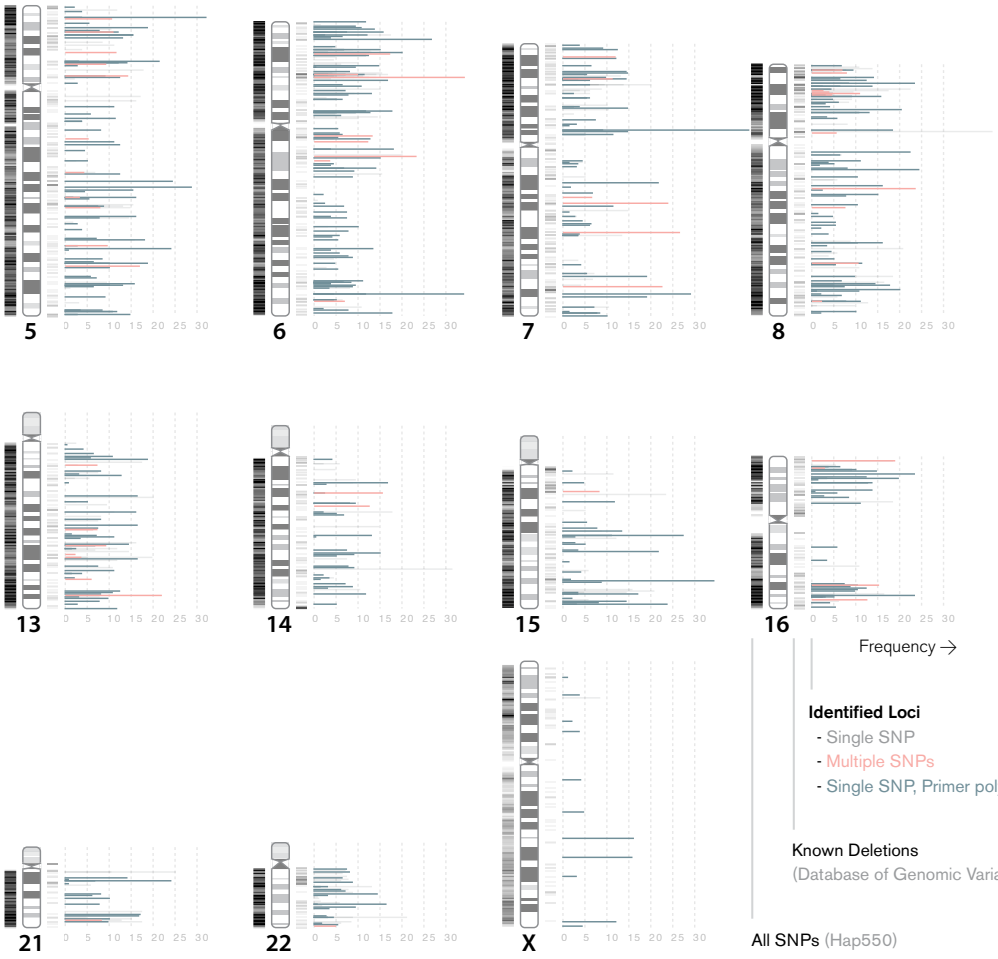

5
6
7
8
13
14
15
16
21
22
X
0 5 10 15 20 25 30
Frequency →
Identified Loci
- Single SNP
- Multiple SNPs
- Single SNP, Primer polymorphism
Known Deletions
(Database of Genomic Variants)
All SNPs (Hap550)
180,000,000 bp

Table 2 Overview of the genomic properties of identified triallelic SNPs

Initial dataset	3,102 samples: 2,185 UK (1,417 on Hap550; 768 on Hap300), 917 Dutch (Hap300)			
Identified triallelic SNPs	1,880			
Identified loci	1,655 (Immediate adjacent triallelic SNPs have been concatenated in a single locus)			
Locus size	**1 triallelic SNP (1,510 Loci)**		**Loci with ≥2 adjacent triallelic SNPs (145 loci)**	
	No known primer polymorphism	Known primer polymorphism	Multiple triallelic SNPs within locus	
	Possible origin of null-allele: - Primer polymorphism - Deletion - Extra allele	Origin of null-allele: - Primer polymorphism	Probable origin of null-allele: - Deletion	
Number of loci	**1073**	**437**	**145**	
Overlap with known CNV deletions	(Enrichment $P<10^{-50}$) 136	50	(Enrichment $P < 10^{-50}$)	77
No. of unique Ensembl genes	490	216	(Enrichment $P = 0.035$)	105
No. of loci that contain Ensembl genes	485	207	(Depletion $P = 0.013$)	59
Median nr of interactions of Ensembl genes	1	1	(Depletion $P = 0.004$)	0
No. of genes with paralogs	359	140	(Enrichment $P = 0.006$)	84
No. of OMIM disease genes	63	30		10
Enriched cytogene-tic bands ($P < 0.05$)	2q, 3p, 6p	6p, 8p, 22q		5p, 8p

nine triallelic SNPs, no primer polymorphism was identified. Additionally, for the samples for which we had predicted a homozygote deletion, no product was observed, suggesting these reflect deletions that are bigger than the loci we had amplified. These results support our estimate that approximately 50% of the triallelic SNPs represent deletions. We also assessed how well the predicted genotypes correspond to the resequenced genotypes. Seventeen SNPs showed perfect concordance, while for six SNPs this was not the case. However, for each of these SNPs, the predicted quality of genotype inference (based on the concordance between the preliminary triallelic genotypes and imputed genotypes) was lower than 0.90, suggesting that genotypes are usually well inferred for triallelic SNPs that have a concordance value over 0.90 (table 3, indicated by the black horizontal bar).

Genomic properties

To gain insight into the enrichment or depletion of certain genomic features within these loci, we analyzed the three triallelic locus categories separately (table 2, if enrichments and depletions P value was below 0.05, these are indicated). Fewer multiple-SNP loci than expected contained genes (Empiric P value = 0.013), but when the loci contained genes, the number of genes was higher than expected (Empiric P value = 0.035). No depletion or enrichment for these measures was observed in the two other classes of loci. It has been demonstrated that genes within CNVs have more paralogs than expected[54]. We also observed this for the multiple SNP loci (Empiric P = 0.006), but not for the other two loci classes. As genes within known deletions tend to be buffered by paralogs that usually have quite similar functions, it is likely that genes within these CNVs are biologically less important. To assess this in a different way, we investigated the number of known interactions these genes have, as various studies have shown[36; 55; 56] that essential genes tend to have more interactions than non-essential genes. We assessed this by analyzing a collection of 80,350 known biological interactions (see Material and methods section), and indeed observed for the genes within the multiple SNP loci that the number of interactions they have is usually significantly less than expected (Wilcoxon-Mann-Whitney P value = 0.004). In addition, various cytogenetic arms (2q, 3p, 5p, 6p, 8p, and 22q) were enriched for triallelic loci (Empiric P value < 0.05).

Summary statistics for the 1,880 triallelic SNPs are provided as supplementary data. *TriTyper* is freely available for downloading from the author's website, along with Java source code. It provides functionality for discovering triallelic SNPs in datasets where raw intensity data is available. When only biallelic genotype calls are available, *TriTyper* allows for imputing triallelic genotypes for 1,204 triallelic SNPs of the 1,880 SNPs we have identified in this study. After assigning triallelic genotypes, *TriTyper* can perform association analysis.

Discussion

In this paper, we have described a method (*TriTyper*) that uses raw intensity data from the Illumina genotyping platform to identify SNPs with an extra untyped, but common allele. Our method is the first to our knowledge to do this in case-control datasets by utilizing the pretablesence of local LD to improve genotype assignments. Through this approach we identified 1,880 triallelic SNPs, and for 1,204 of these the LD patterns permitted inferring the triallelic genotypes without needing access to raw intensity data. This enables association analyses on these SNPs in Caucasian datasets that have similar LD patterns, but for which only genotype calls have been made available, or those that have been generated using completely different platforms.

By using the triallelic genotype calls from *TriTyper,* highly robust association analyses can be performed. We have shown this in a triallelic null-allele association analysis in celiac disease, where cases had been run on a different type of array than that used for the controls, and we saw no inflation of the test statistic. Simulations indicate that our method has superior power to detect these associations, compared to an association analysis on the biallelic SNPs that are in LD and have been used to infer the triallelic genotypes. The triallelic SNPs identified

Table 3 Resequencing results of triallelic SNPs

SNP	Known primer polymorphism	Genotyped samples (predicted inferred genotypes)		Predicted Reliability Statistic
rs10504729	-	6	(1 A0, 1 AA, 1 AG, 2 G0, 1 GG)	0.66
rs2675899 *	-	10	(2 A0, 2 AA, 2 CA, 2 C0, 2 CC)	0.71
rs13213842	rs35678510, A/G, +1 bp	8	(1 A0, 1 AA, 2 AG, 2 G0, 2 GG)	0.71
rs3131755	-	5	(1 A0, 1 AA, 1 B, 2 BB)	0.75
rs195738	-	12	(1 00, 4 A0, 1 AA, 4 G0, 1 GG)	0.79
rs8053391 *	-	4	(2 A0, 1 AA, 1 GG)	0.81
rs7678151	rs28542567, A/G, -3 bp	11	(1 00, 2 A0, 2 AA, 2 AG, 2 G0, 2 GG)	0.83
rs2871198 *	-	10	(4 A0, 1 AA, 1 AG, 3 G0, 1 GG)	0.86
rs9355606	-	6	(1 A0, 1 AG, 2 G0, 2 GG)	0.86
rs495991	-	4	(2 G0, 2 AG)	0.86
rs10510312	-	6	(3 G0, 1 GG, 2 AG)	0.87
rs2486674	-	18	(9 A0, 1 AA, 1 AG, 6 G0, 1 GG)	0.88
rs11834116	-	4	(1 AA, 1 A0, 1 AG, 1 G0)	0.91
rs7083969	-	9	(1 A0, 1 AA, 1 AG, 5 G0, 1 GG)	0.92
rs7083969	-	11	(6 A0, 1 AA, 1 AG, 3 G0, 1 GG)	0.92
rs1109374	-	4	(1 A0, 1 AA, 1 AG, 1 GG)	0.92
rs9361448	-	14	(1 00, 5 A0, 1 AA, 1 AC, 5 C0, 1 CC)	0.94
rs11533655 *	-	15	(2 00, 5 A0, 1 AA, 1 AG, 5 G0, 1 GG)	0.95
rs2254039	-	6	(3 G0, 2 GG, 1 AG)	0.95
rs2894386 *	-	17	(2 A0, 8 AA)	0.95
rs7133541	-	10	(5 A0, 1 AA, 1 AG, 2 G0, 1 GG)	0.96
rs7822381	-	18	(4 A0, 7 AA, 5 AG, 2 G0)	0.97
rs1551821	-	6	(3 A0, 2 AC, 1 AA)	0.98

* Known deletion locus (Database of Genomic Variants, March 2007 release)

Observed origin of null allele	Resequenced discordant genotypes (explanation)	
Primer polymorphism (C/T, -1 bp)	0	
Probably deletion	0	
Primer polymorphism (A/G, +1 bp)	1	(G0 > GG)
Primer polymorphism (T/G, +4 bp)	0	
Probably deletion	2	(A0 > AA)
Primer polymorphism (C/G, +4 bp)	0	
Primer polymorphism (A/G, -3 bp)	1	(G0 > GG)
Probably deletionw	0	
Probably deletion	0	
Primer polymorphism (C/G, -1 bp)	1	(G0 > GG)
Primer polymorphism (A/C, -1 bp)	1	(G0 > GG)
Deletion (TGAGTATAGTAdel>AGTTTins/+)	5	(3 A0 > AA, 2 G0 > GG)
Primer polymorphisms (C/T, -8 bp, A/G, + 1bp)	0	
Probably deletion	0	
Probably deletion	0	
Primer polymorphism (C/T, +3 bp)	0	
Probably deletion	0	
Probably deletion	0	
Primer polymorphism (C/T, -1 bp)	0	
Primer polymorphism (C/T, -4 bp)	0	
Probably deletion	0	
Deletion (1bp deletion in primer)	0	
Primer polymorphism (A/C, +1 bp)	0	

Total - 2 known primer polymorphisms
- 10 previously unknown primer polymorphisms
- 11 probably deletions

Figure 4

Distribution of triallelic locus size and null-allele frequency

a) The triallelic loci for which no polymorphism within the primer are known in dbSNP (build 127) are plotted against the maximum potential size of each locus, assuming these can reflect deletions (by taking the physical position of the immediately adjacent SNPs that look normal on the Illumina Human Hap550 platform). **b)** The number of triallelic loci is plotted against the null-allele frequency.

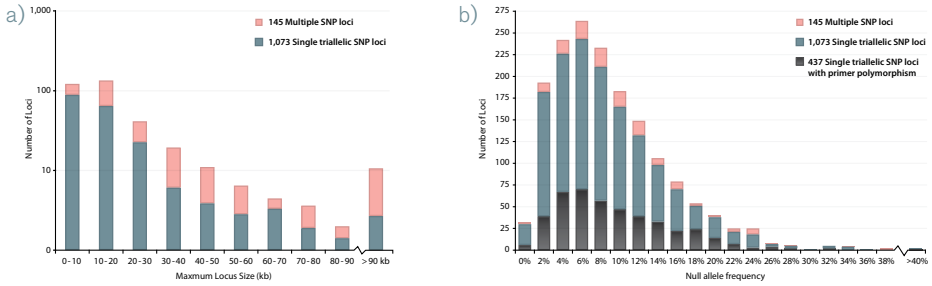


Figure 5

Consequences of mistyping a null-allele for case-control association studies

It is assumed allele A is the true risk-allele for various values of γ (relative risk of AA-homozygote) and frequencies of the null-allele (p_0). The overestimation of the effect under a dominant model (top figure) and the underestimation of the effect under a recessive model (bottom figure) are shown.

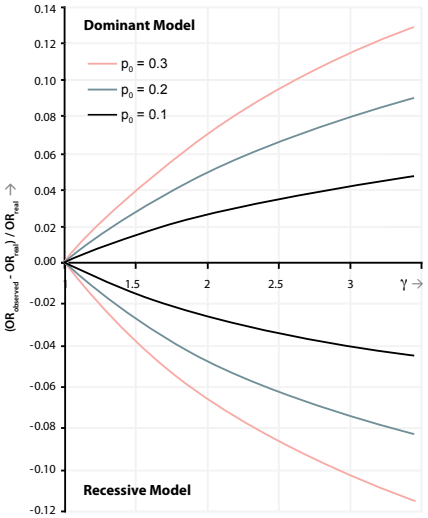
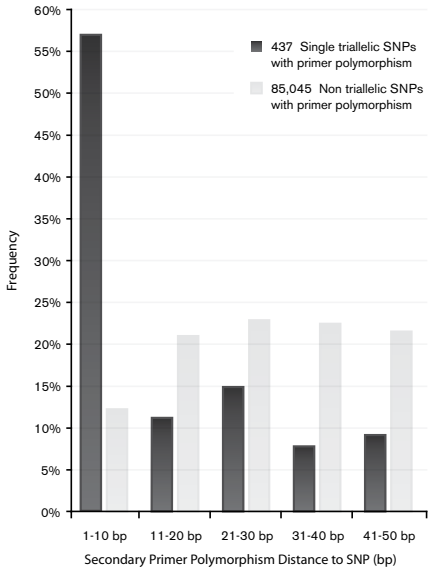


Figure 6

Distribution of distance of secondary polymorphisms present within primers of Human Hap550 SNPs

Distribution plot of the distance of secondary polymorphisms present within primers of Human Hap550 SNPs to the actual SNP. Polymorphisms are known in dbSNP (build 127) for 85,045 of the SNPs present on the Illumina Hap550 platform within the 50 bp long primers. For the 1,880 fitted triallelic SNPs, this is the case for 437 of the SNPs (Expected 235, Fisher's Exact P value $< 10^{-18}$). When investigating how far away these secondary polymorphisms are from the actual SNPs, it turns out that within the triallelic SNPs these secondary polymorphisms usually map much closer to the actual SNP than for the non-triallelic SNPs (Wilcoxon-Mann-Whitney P value $< 10^{-76}$).



also have ramifications for association analyses that are based on biallelic assumptions. If, for any of the triallelic SNPs, the null-allele is not associated but the A and B alleles are, the real effect of the association will be over- or underestimated, depending on a dominant or recessive model, respectively.

The reported associations in celiac disease did not survive multiple testing when we assumed hundreds of thousands of biallelic association tests have already been performed in a GWA analysis. These findings, however, do provide new hypotheses for further replication in independent cohorts.

The identity of each of the triallelic SNPs identified remains to be established. We observed that 437 triallelic SNPs showed a triallelic pattern because of a polymorphism in the region of the primer, usually within 10 bp from the target SNP (see figure 6). This artifact should serve as a warning for all oligonucleotide-based assays and we urge researchers to validate putative CNVs with different techniques. For the remaining 1,218 unique loci (where immediately adjacent triallelic SNPs had been concatenated), we observed a strong enrichment for deletions, known in the Database of Genomic Variants. We estimate that, of these loci, 682 reflect deletions, suggesting that on average 99 deletion CNVs per individual were identified. This is approximately four times more than what has been found by other methods using identical oligonucleotide arrays (between 10 and 27 CNVs on average per individual[1; 14; 22]). The high-resolution of our method and the fact that we take LD into account probably explain this difference.

Loci that contained multiple SNPs overlapped with fewer genes than expected, although the total number of genes for these loci was higher than expected. Comparable analyses[1; 54] conflict with each other, which warrants further clarification. As shown before[54], genes within these loci have paralogs more often than expected (P value = 0.006). We are the first to our knowledge to show that the genes within these loci also biologically interact with significantly fewer genes than expected (P value = 0.004).

Various avenues for extending *TriTyper* can be envisaged. A drawback of our current imputation methodology is that we assume certain haplotypes have a zero frequency, which might not reflect the reality due to lower LD than assumed. Therefore, for some of the triallelic SNPs it is likely some of the imputed genotypes will be incorrect. Consequently, an association analysis using imputed triallelic genotypes will have lower statistical power compared to an ideal situation, where accurate triallelic genotypes would be available. We argue this sacrifice in calling accuracy and power due to imputation is acceptable, since it considerably reduces Type I errors in association testing. If different platforms or batches have been used for genotyping and cases and controls are not evenly spread[28] over these, spurious associations are to be expected due to the way our calling algorithm initially discriminates between A0 and AA, and B0 and BB genotypes. If these genotypes can be imputed using nearby biallelic SNPs, false-positive associations will be prevented. Although highly sophisticated imputation algorithms have been described for biallelic SNPs[26; 57], it is not straightforward to use these to resolve this issue. This is mostly due to the fact that we currently cannot rely upon phased haplotypes from HapMap, because all the SNPs within HapMap have been called under biallelic assumptions. Another complication is the difficulty to estimate r^2 and to interpret D' if the number of alleles between two markers differ[58; 59]. However, we expect that by incorporating some of the concepts underlying these biallelic imputation methodologies, the accuracy of the imputed triallelic genotypes can be improved.

Currently, *TriTyper* can only detect SNPs with a common extra but untyped allele. We envisage that adaptations to both our calling algorithm and LD-based genotype imputation methodology will probably allow identification of very small but common duplications. In addition, studies that aim to identify rare *de novo* deletions and duplications can immediately benefit from our work. As the number of samples we have studied is reasonably high (3,102), we were able to identify common triallelic SNPs that had a null-allele frequency as low as 0.5%. If

researchers are not aware of these common triallelic SNPs and use smaller cohorts, they might deem these SNPs rare and potentially biologically interesting when aberrant characteristics are observed in only a few samples. Methodologically, the resolution of *de novo* CNV detection methods[14;22] can also be improved by incorporating LD-based frameworks: conceptually, if two SNPs are in very strong LD, but in one sample a recombination seems to be present, a *de novo* duplication or deletion that spans one of these SNPs could be an alternative explanation.

The Illumina BeadChip arrays we have used here are strongly biased against CNVs, because SNPs that showed low call rates, HWE deviations or many Mendelian segregation inconsistencies in a subset of the HapMap samples had been removed during the design of these chips. This also explains why the observed median null-allele frequency of the identified triallelic SNPs was only 7.6%. Since we did not use the most current Illumina chips, we expect the newer ones that are better tailored to target CNVs (e.g. Illumina HumanHap370 and HumanHap1M), to lead to greater insight into CNVs.

The Human Gene Mutation Database[60] reports 73,411 variants that mostly have a phenotypic effect, of which about 16% are micro-deletions and 7% are micro-insertions (smaller than 20 bp), whereas larger deletions and insertions constitute 6% and 1% of the variants, respectively. This clearly indicates the importance of structural variants and deletions in both rare and common diseases[6-8]. New statistical CNV detection methods (like *TriTyper*) and more extensive oligonucleotide arrays will undoubtedly result in the identification of many more variants, of which quite a few will turn out to be associated with disease.

Acknowledgments

We thank Jackie Senior, Madelien van de Beek, Ritsert Jansen and members of the Complex Genetics Section, UMC Utrecht for critically reading the manuscript. We thank D. Simpkin, T. Dibling and C. Hand for genotyping (Sanger Institute); and D. Strachan and W.L. McArdle for 1958 Birth Cohort samples. We thank Illumina for providing HapMap genotype data. We thank Dutch and UK clinicians who collected samples[28;30], and sample donors. We thank the Genomics Center Utrecht for computational resources. Statistical analyses were carried out on the Genetic Cluster Computer in Amsterdam, which is financially supported by the Netherlands Organization for Scientific Organization (NWO, grant 480-05-003). We acknowledge funding from Coeliac UK; the Netherlands Organization for Scientific Research (NWO, grant 918-66-620); Netherlands Organization for Health Research and Development (ZonMW grant 917-66-315); the Coeliac Disease Consortium (an innovative cluster approved by the Netherlands Genomics Initiative and partly funded by the Dutch government (grant BSIK03009)); the Netherlands Genomics Initiative (grant 050-72-425 and fellowship grant to L.F.); Prinses Beatrix Fonds (L.H.v.d.B.); the Wellcome Trust (GR068094MA Clinician Scientist Fellowship to D.A.v.H.; and support for the work of P.D.). The authors acknowledge use of genotypes from the British 1958 Birth Cohort collection, funded by the UK Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02.

Web Resources

The URL for data presented here is as follows:

TriTyper www.ludesign.nl/trityper

References

1 Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, *et al.* (2006) Global variation in copy number in the human genome. Nature 444:444-454

2 McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, Altshuler DM (2006) Common deletion polymorphisms in the human genome. Nature genetics 38:86-92

3 Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C (2004) Detection of large-scale variation in the human genome. Nature genetics 36:949-951

4 Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK (2006) A high-resolution survey of deletion polymorphism in the human genome. Nature genetics 38:75-81

5 de Smith AJ, Tsalenko A, Sampas N, Scheffer A, Yamada NA, Tsang P, Ben-Dor A, Yakhini Z, Ellis RJ, Bruhn L, Laderman S, Froguel P, Blakemore AI (2007) Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. Human molecular genetics 16:2783-2794

6 Aitman TJ, Dong R, Vyse TJ, Norsworthy PJ, Johnson MD, Smith J, Mangion J, Roberton-Lowe C, Marshall AJ, Petretto E, Hodges MD, Bhangal G, Patel SG, Sheehan-Rooney K, Duda M, Cook PR, Evans DJ, Domin J, Flint J, Boyle JJ, Pusey CD, Cook HT (2006) Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. Nature 439:851-855

7 Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, Murthy KK, Rovin BH, Bradley W, Clark RA, Anderson SA, O'Connell R J, Agan BK, Ahuja SS, Bologna R, Sen L, Dolan MJ, Ahuja SK (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science (New York, NY 307:1434-1440

8 Fellermann K, Stange DE, Schaeffeler E, Schmalzl H, Wehkamp J, Bevins CL, Reinisch W, Teml A, Schwab M, Lichter P, Radlwimmer B, Stange EF (2006) A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. American journal of human genetics 79:439-448

9 Wong KK, deLeeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, Horsman DE, MacAulay C, Ng RT, Brown CJ, Eichler EE, Lam WL (2007) A comprehensive analysis of common copy-number variations in the human genome. American journal of human genetics 80:91-104

10 Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D, Eichler EE (2005) Segmental duplications and copy-number variation in the human genome. American journal of human genetics 77:78-88

11 Simon-Sanchez J, Scholz S, Fung HC, Matarin M, Hernandez D, Gibbs JR, Britton A, de Vrieze FW, Peckham E, Gwinn-Hardy K, Crawley A, Keen JC, Nash J, Borgaonkar D, Hardy J, Singleton A (2007) Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. Human molecular genetics 16:1-14

12 Pinto D, Marshall C, Feuk L, Scherer SW (2007) Copy-number variation in control population cohorts. Human molecular genetics 16 Spec No. 2:R168-173

13 Komura D, Shen F, Ishikawa S, Fitch KR, Chen W, Zhang J, Liu G, Ihara S, Nakamura H, Hurles ME, Lee C, Scherer SW, Jones KW, Shapero MH, Huang J, Aburatani H (2006) Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. Genome research 16:1575-1584

14 Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J (2007) QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic acids research 35:2013-2025

15 Kohler JR, Cutler DJ (2007) Simultaneous discovery and testing of deletions for disease association in SNP genotyping studies. American journal of human genetics 81:684-699

16 Kosta K, Sabroe I, Goke J, Nibbs RJ, Tsanakas J, Whyte MK, Teare MD (2007) A Bayesian approach to copy-number-polymorphism analysis in nuclear pedigrees. American journal of human genetics 81:808-812

17 Nannya Y, Sanada M, Nakazaki K, Hosoya N, Wang L, Hangaishi A, Kurokawa M, Chiba S, Bailey DK, Kennedy GC, Ogawa S (2005) A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer research 65:6071-6079

18 Zhang J, Feuk L, Duggan GE, Khaja R, Scherer SW (2006) Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. Cytogenetic and genome research 115:205-214

19 Leykin I, Hao K, Cheng J, Meyer N, Pollak MR, Smith RJ, Wong WH, Rosenow C, Li C (2005) Comparative linkage analysis and visualization of high-density oligonucleotide SNP array data. BMC genetics 6:7

20 Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851-861

21 Carter NP (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. Nature genetics 39:S16-21

22 Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome research 17:1665-1674

23 Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS (2005) A genome-wide scalable SNP genotyping assay using microarray technology. Nature genetics 37:549-554

24 Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA (2006) Common deletions and SNPs are in linkage disequilibrium in the human genome. Nature genetics 38:82-85

25 Locke DP, Sharp AJ, McCarroll SA, McGrath SD, Newman TL, Cheng Z, Schwartz S, Albertson DG, Pinkel D, Altshuler DM, Eichler EE (2006) Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. American journal of human genetics 79:275-290

26 Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nature genetics 39:906-913

27 Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, *et al.* (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science (New York, NY 316:1341-1345

28 van Heel DA, Franke L, Hunt KA, Gwilliam R, Zhernakova A, Inouye M, Wapenaar MC, *et al.* (2007) A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nature genetics 39:827-829

29 Ceppellini R, Siniscalco M, Smith CA (1955) The estimation of gene frequencies in a random-mating population. Annals of human genetics 20:97-115

30 van Es MA, van Vught PW, Blauw HM, Franke L, Saris CG, Van den Bosch L, de Jong SW, de Jong V, Baas F, van't Slot R, Lemmens R, Schelhaas HJ, Birve A, Sleegers K, Van Broeckhoven C, Schymick JC, Traynor BJ, Wokke JH, Wijmenga C, Robberecht W, Andersen PM, Veldink JH, Ophoff RA, van den Berg LH (2008) Genetic variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis. Nature genetics 40:29-31

31 Carlson CS, Smith JD, Stanaway IB, Rieder MJ, Nickerson DA (2006) Direct detection of null alleles in SNP genotyping data. Human molecular genetics 15:1931-1937

32 Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, *et al.* (2007) Ensembl 2007. Nucleic acids research 35:D610-617

33 Stephensen AG (2002) EVD: extreme value distributions. R-News:31-32

34 McKusick VA (2007) Mendelian Inheritance in Man and its online version, OMIM. American journal of human genetics 80:588-604

35 Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic acids research 32:D277-280

36 Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP, Vidal M (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 430:88-93

37 Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, Matthews L, Wu G, Birney E, Stein L (2007) Reactome: a knowledge base of biologic pathways and processes. Genome biology 8:R39

38 Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, *et al.* (2005) The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic acids research 33:D418-424

39 Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, *et al.* (2006) Human protein reference database--2006 update. Nucleic acids research 34:D411-414

40 Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H (2007) IntAct--open source resource for molecular interaction data. Nucleic acids research 35:D561-565

41 Schymick JC, Scholz SW, Fung HC, Britton A, Arepalli S, Gibbs JR, Lombardo F, Matarin M, Kasperaviciute D, Hernandez DG, Crews C, Bruijn L, Rothstein J, Mora G, Restagno G, Chio A, Singleton A, Hardy J, Traynor BJ (2007) Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. Lancet neurology 6:322-328

42 Fung HC, Scholz S, Matarin M, Simon-Sanchez J, Hernandez D, Britton A, Gibbs JR, Langefeld C, Stiegert ML, Schymick J, Okun MS, Mandel RJ, Fernandez HH, Foote KD, Rodriguez RL, Peckham E, De Vrieze FW, Gwinn-Hardy K, Hardy JA, Singleton A (2006) Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. Lancet neurology 5:911-916

43 Sollid LM (2000) Molecular basis of celiac disease. Annual review of immunology 18:53-81

44 Karell K, Louka AS, Moodie SJ, Ascher H, Clot F, Greco L, Ciclitira PJ, Sollid LM, Partanen J (2003) HLA types in celiac disease patients not carrying the DQA1*05-DQB1*02 (DQ2) heterodimer: results from the European Genetics Cluster on Celiac Disease. Human immunology 64:469-477

45 Monsuur AJ, de Bakker PI, Alizadeh BZ, Zhernakova A, Bevova MR, Strengman E, Franke L, van't Slot R, van Belzen MJ, Lavrijsen IC, Diosdado B, Daly MJ, Mulder CJ, Mearin ML, Meijer JW, Meijer GA, van Oort E, Wapenaar MC, Koeleman BP, Wijmenga C (2005) Myosin IXB variant increases the risk of celiac disease and points toward a primary intestinal barrier defect. Nature genetics 37:1341-1344

46 Hunt KA, Zhernakova A, Turner G, Heap G, Franke L, Bruinenberg M, Romanos J, *et al.* (2008) Novel coeliac disease genetic risk loci with links to adaptive immunity. Nature genetics In Press

47 Liu J, Juo SH, Holopainen P, Terwilliger J, Tong X, Grunn A, Brito M, Green P, Mustalahti K, Maki M, Gilliam TC, Partanen J (2002) Genomewide linkage analysis of celiac disease in Finnish families. American journal of human genetics 70:51-59

48 Greco L, Babron MC, Corazza GR, Percopo S, Sica R, Clot F, Fulchignoni-Lataud MC, Zavattari P, Momigliano-Richiardi P, Casari G, Gasparini P, Tosi R, Mantovani V, De Virgiliis S, Iacono G, D'Alfonso A, Selinger-Leneman H, Lemainque A, Serre JL, Clerget-Darpoux F (2001) Existence of a genetic risk factor on chromosome 5q in Italian coeliac disease families. Annals of human genetics 65:35-41

49 Greco L, Corazza G, Babron MC, Clot F, Fulchignoni-Lataud MC, Percopo S, Zavattari P, *et al.* (1998) Genome search in celiac disease. American journal of human genetics 62:669-675

50 Babron MC, Nilsson S, Adamovic S, Naluai AT, Wahlstrom J, Ascher H, Ciclitira PJ, Sollid LM, Partanen J, Greco L, Clerget-Darpoux F (2003) Meta and pooled analysis of European coeliac disease data. Eur J Hum Genet 11:828-834

51 Riccioni R, Saulle E, Militi S, Sposi NM, Gualtiero M, Mauro N, Mancini M, Diverio D, Lo Coco F, Peschle C, Testa U (2003) C-fms expression correlates with monocytic differentiation in PML-RAR alpha+ acute promyelocytic leukemia. Leukemia 17:98-113

52 Zapata-Velandia A, Ng SS, Brennan RF, Simonsen NR, Gastanaduy M, Zabaleta J, Lentz JJ, Craver RD, Correa H, Delgado A, Pitts AL, Himel JR, Udall JN, Jr., Schmidt-Sommerfeld E, Brown RF, Athas GB, Keats BB, Mannick EE (2004) Association of the T allele of an intronic single nucleotide polymorphism in the colony stimulating factor 1 receptor with Crohn's disease: a case-control study. J Immune Based Ther Vaccines 2:6

53 Estivill X, Armengol L (2007) Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. PLoS genetics 3:1787-1799

54 Nguyen DQ, Webber C, Ponting CP (2006) Bias of selection on human copy-number variants. PLoS genetics 2:e20

55 Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411:41-42

56 Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL (2007) The human disease network. Proceedings of the National Academy of Sciences of the United States of America 104:8685-8690

57 Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. American journal of human genetics 78:629-644

58 Hedrick PW (1987) Gametic disequilibrium measures: proceed with caution. Genetics 117:331-341

59 Zapata C (2000) The D' measure of overall gametic disequilibrium between pairs of multiallelic loci. Evolution 54:1809-1812

60 Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN (2003) Human Gene Mutation Database (HGMD): 2003 update. Human mutation 21:577-581

61 Yu Z, Schaid DJ (2007) Methods to impute missing genotypes for population data. Hum Genet

62 Slatkin M, Excoffier L (1996) Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. Heredity 76 (Pt 4):377-383

Appendix A

Conventional biallelic genotype-calling

When the minor allele frequency (MAF) is sufficiently high, assigning genotypes to biallelic SNPs is usually fairly straightforward: three separate clusters will appear (reflecting the AA, AB and BB genotypes) that can usually be well separated using a clustering algorithm we recently described[28]. This algorithm uses per sample polar angle θ ($\theta = 2/\pi$ * arctan (intensity$_b$ / intensity$_a$)) to identify three clusters of sample for which the standard deviations of the θ values for each cluster are low. This is achieved by exploring a two-dimensional search space (where one parameter discriminates between AA and AB samples and the other discriminates between AB and BB samples). The method then settles upon a certain clustering for which the three calculated standard deviations have a sum that has been minimized.

Preliminary triallelic genotype-calling

When a SNP is triallelic, but the SNP has been called under biallelic assumptions for sufficient samples, it is likely that HWE deviations will be observed. Assuming HWE for the true alleles A, B and 0, we can compute the expected frequencies of *observed* genotypes AA, AB and BB. From these we can compute the *observed* allele frequencies for A and B. Now the deviation from the Hardy Weinberg equilibrium in those observed genotypes AA, AB and BB, relative to the genotype frequencies expected from the observed allele frequencies A and B can be computed. It turns out that the resulting χ^2 depends on the true frequency of the 0 allele, and of course on the sample size, but not on the frequencies of the A and B alleles:

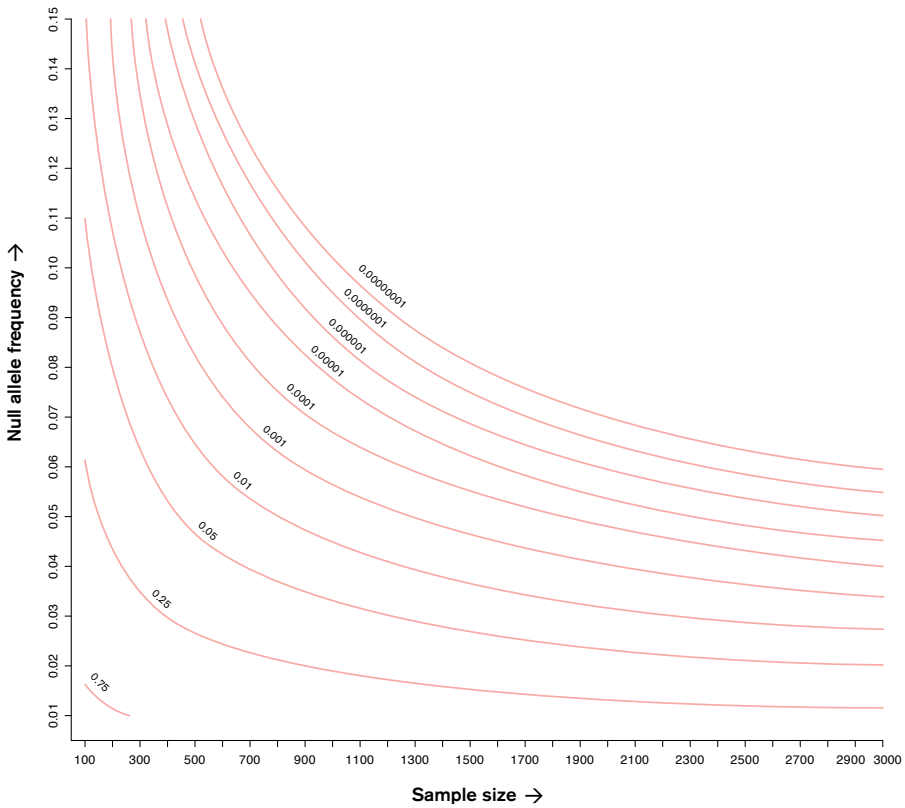
$$\chi^2 = n \cdot p_0^2 \cdot \left(4 - 8p_0 + 5p_0^2\right)$$

Where n is sample size and p_0 the frequency of the 0 allele.

Figure 7

HWE test statistics, when analyzing triallelic SNPs, called under biallelic assumptions

These calculations show the HWE test statistic P value for various sample sizes and different frequencies of the 0 allele. If we incorrectly assume triallelic SNPs are biallelic, analysis of sample sizes that are representative of current genome-wide association studies will result in significant HWE deviations, even when the 0 allele has a fairly low frequency, *e.g.* when testing 3,000 samples, and assuming a call rate of 100% for samples having 1 or 2 copies of the A or B allele, triallelic SNPs with a null-allele frequency above 2% have an expected deviation from HWE with P value < 0.05.



Calculations show that if 3,000 samples are typed, a null-allele with a frequency of 2% or higher will on average cause a deviation from HWE that can be demonstrated at the level of P = 0.05. Figure 7 illustrates how the HWE test statistic depends on the sample size and the frequency of the 0 allele.

Although these HWE deviations can also arise due to failed assays, they are explained by an unlabelled allele in a substantial number of cases[31]. We followed up SNPs when, under biallelic assumptions, the exact HWE P value was below 0.05 or when the call rate was below 98%. For these SNPs we determined whether triallelic genotypes could be called by introducing two additional parameters (α and β) to our calling algorithm.

In the initial triallelic genotype-calling procedure, genotypes 00 are assigned to samples that have a Euclidian intensity below α. For the remaining samples we use the aforementioned calling algorithm to identify three clusters of samples that are either A0 or AA (A0/AA), are AB, or either are B0 or BB (B0/BB) (fig. 1B).

Subsequently we partition both the A0 and AA samples and the B0 and BB samples using parameter β. Non-pseudoautosomal chromosome X SNPs provide detailed insight into the intensity characteristics of these A0, B0, AA and BB samples. For these SNPs, females will usually have two copies, whereas males will only have one copy (fig. 8A). We investigated 11,652 non-pseudoautosomal chromosome X SNPs, present on the Illumina Human Hap550 platform, for which 1,417 unrelated UK samples from the 1958 British Birth Cohort had been typed[28]. For each of these SNPs we linearly scaled the probe intensities, such that the center of the AB cluster was at coordinate (1, 1). We then moved the origin of the Cartesian coordinate system to this coordinate and converted to a polar coordinate system, allowing us to determine a one-dimensional angle distribution for both the A0, the AA, the B0 and BB samples. These distributions allow us to introduce parameter β (range [0, 100]), which denotes both the percentile of the A0 and the percentile of the B0 distributions. We use this parameter to distinguish between one and two copies (fig. 1C) as the corresponding percentile corresponds to two different Cartesian rays that both start from the AB cluster center but have different angles, where one ray (reflecting the percentile within the chromosome X A0 distribution) allows us to divide the A0/AA samples in A0 and AA samples and another ray (reflecting the percentile within the chromosome X B0 distribution) allows us to divide the B0/BB samples in B0 and BB samples (fig. 8B). *E.g.* when $\beta = 25$ (fig. 8B, left), for the samples which are either AA or A0, the samples having an angle to the AB cluster location below 260° will be designated A0, and above 260° will be designated AA. For samples that are either BB or B0, those having an angle to the AB cluster location below 192° will be designated B0. When β=75 (fig. 8b, right), the thresholds for these angles are 271° and 184°, respectively.

It is evident that different α and β values will result in different triallelic genotype assignments. To optimize these we use an MLE procedure that assumes HWE under a triallelic model, through the following log likelihood formula[29]:

$$\log(\text{likelihood}) = \log\left[(n_{aa} + n_{bb} + n_{ab} + n_{a0} + n_{b0} + n_{00})!\right] - \left[\log(n_{aa}!) + \log(n_{bb}!) + \log(n_{ab}!) + \log(n_{a0}!) + \log(n_{b0}!) + \log(n_{00}!)\right] + n_{aa} * \log(p_a * p_a) + n_{ab} * \log(2 * p_a * p_b) + n_{bb} * \log(p_b * p_b) + n_{a0} * \log(2 * p_a * p_0) + n_{b0} * \log(2 * p_b * p_0) + n_{00} * \log(p_0 * p_0)$$

where n_{aa}, n_{bb}, n_{ab}, n_{a0}, n_{b0} and n_{00} are the number of individuals with assigned genotype AA, BB, AB, A0, B0 and 00, respectively, and p_a, p_b and p_0 are the allele frequencies of allele A, B and 0, respectively.

Through analysis of the entire search space, the values for α and β for which this likelihood is maximal can be determined (fig. 1D), indicating that the assigned genotype distribution most closely resembles the distribution expected under triallelic HWE. Identified triallelic SNPs are included for follow-up analysis, if the null-allele frequency is over 0.5% and the fitted β parameter value is between 6 and 97.

Eventual triallelic genotype-calling through imputation

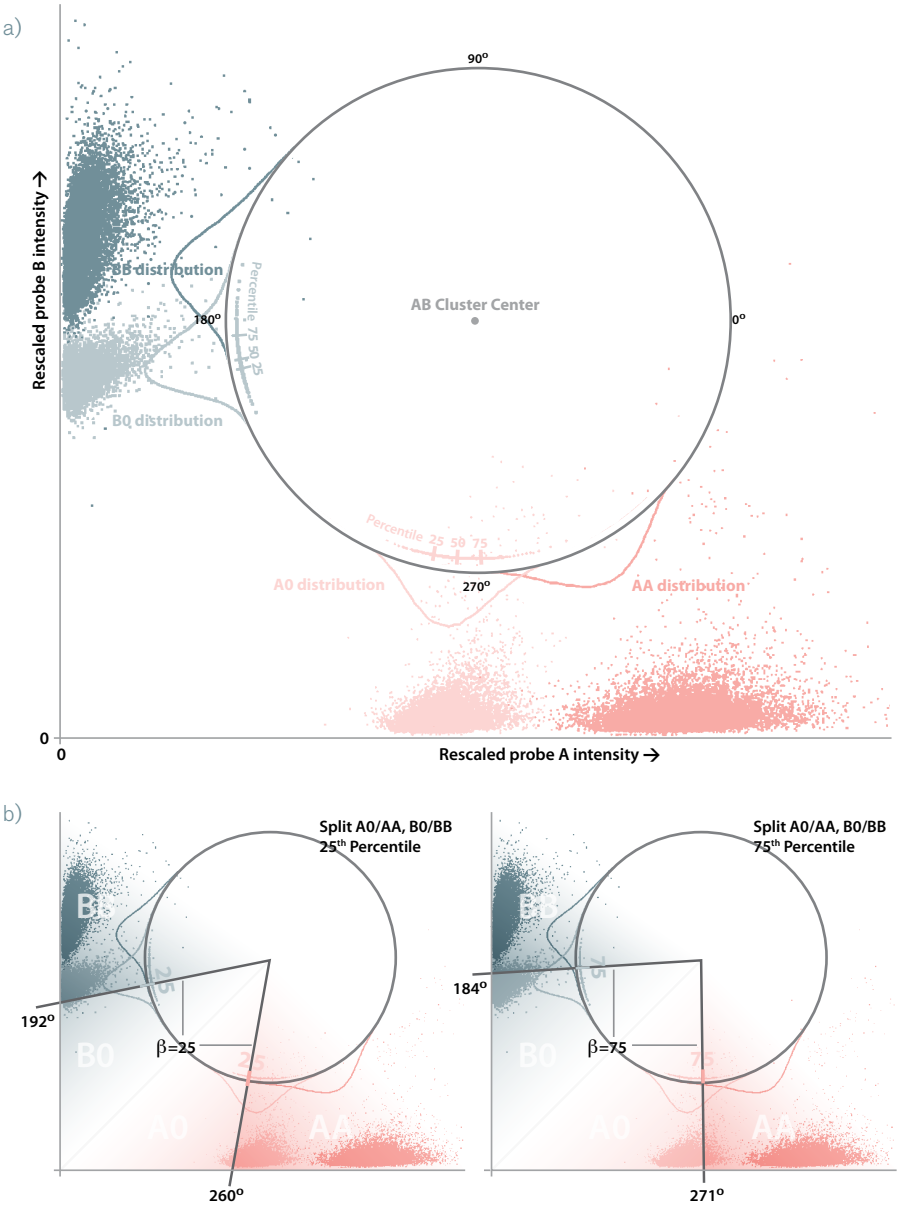
To improve upon the initially assigned triallelic genotypes, we take advantage of local linkage disequilibrium, because the presence of LD between biallelic SNPs can often be utilized to improve genotype assignments[26;27;57]. As LD has been described for deletion CNVs as well[1;2;24;25;61], we assumed these triallelic genotypes can potentially also be inferred through LD.

To assess this we require that at least one of the six haplotypes should have a zero frequency and that all alleles are present for the biallelic SNP and triallelic SNP, resulting in the identification of 24 'haplotype scenarios', that each have a different set of haplotypes that have not been observed (Figure 9). For each of these scenarios a set of triallelic genotype imputation rules can be easily deduced. It turns out that ten scenarios are capable of discriminating between A0 and AA and/or between B0 and BB triallelic genotypes. This is very helpful as in the initial genotype assignment procedure a somewhat rough division is made between the A0 and AA genotypes and between the B0 and BB genotypes (through optimization of parameter β). As such, it is likely that some incorrect genotypes (fig. 1E) have initially been assigned to samples that cluster in the vicinity of the two dividing rays determined by parameter β (e.g. the initially assigned A0 genotype should actually be AA and vice versa). This is resolved if nearby biallelic SNPs allow for discrimination between A0 and AA and between B0 and BB samples. We concentrate on any of these ten scenarios throughout this paper and will assess these for each triallelic SNP.

Figure 8

Distribution of A and B allele intensities of 11,652 chromosome X SNPs, present on the Human HapMap550 platform

Each dot represents the median coordinate of either the A0 (males, green), AA (females, blue), B0 (males, yellow), BB (females, red) or AB (females, grey) cluster for a single SNP. The A and B intensities have been scaled in such a way that for each SNP the median AB cluster center is identical for all chromosome X SNPs. **a)** It is evident that single copy genotypes (A0 and B0) clearly show different intensity characteristics than AA and BB genotypes. Another observation is that the A and B probes have slightly different characteristics, as the A0 and AA distributions overlap slightly less than the B0 and BB distributions, indicating that on average A0 and AA samples can be better distinguished from each other. To correct for these differences in intensity characteristics, parameter β is calibrated on these chromosome X SNP distributions. **b)** The genotype-calling algorithm uses parameter b to distinguish between A0 and AA and between B0 and BB. For a given b an angle for the A0 distribution is determined where the A0 distribution percentile equals β. The same holds for the angle of the B0 distribution. In the present example, increasing β increases the angle of the A-line slightly more than it increases the angle of the B-line. Examples are shown where β is 25 (figure b, left) and where β is 75 (figure b, right), resulting in different genotype assignments (A0, AA, B0 and BB genotypes assignments are indicated in 50% red, 100% red, 50% metallic and 100% metallic, respectively).



We first assess the LD for each triallelic SNP identified with the immediately adjacent biallelic SNPs (10 to the left and 10 to the right): for each pair, haplotype frequencies (h_{aa}, h_{ab}, h_{ba}, h_{bb}, h_{0a} and h_{0b}) are estimated using an expectation-maximization algorithm[62]. If the frequencies of some of these haplotypes are zero (e.g. haplotypes h_{aa}, h_{ba} and h_{0b} have a zero frequency, as in figure 1F), it is determined whether this configuration of observed and non-observed haplotypes matches one of the ten haplotype scenarios for which the biallelic SNP helps to discriminate between some of the triallelic genotypes, we use the neighboring SNP for imputation. Due to the uncertainties mentioned for the initially assigned triallelic genotypes, certain estimated haplotypes frequencies will be incorrect, resulting in haplotypes with non-zero frequencies that in reality should have a zero frequency (fig. 1F). In order to overcome this we relaxed our method for assessing the imputation potential of each neighboring biallelic SNP: we assumed that haplotypes with low, but non-zero frequencies, in reality might have a zero frequency. For each haplotype it was determined whether the frequency was lower than the frequency of the haplotype with the same triallelic allele, but with a different biallelic allele. If this was the case, we assumed that this haplotype in reality might have a zero frequency. To ascertain this, we tested all possible haplotype scenarios (through systematic inclusion and exclusion of these potentially zero-frequency haplotypes) and assessed whether any of these scenarios could help to discriminate between A0 and AA or between B0 and BB. If this was observed we searched for evidence that our zero-frequency assumption for these haplotypes was indeed correct, by imputing the A0 and AA or B0 and BB genotypes and testing whether the Euclidian intensities of the imputed A0 or B0 samples were significantly lower (Wilcoxon-Mann-Whitney test $P < 10^{-3}$) than the Euclidian intensities of the AA or BB samples. In addition we tested whether the concordance between the imputed and observed genotypes was higher than 60%. If this was observed we assumed this haplotype scenario could be used for imputation purposes and stored it in a vector. Once all haplotype scenarios had been assessed for each of the twenty biallelic neighboring SNPs, we selected the imputation scenario with the highest genotypic concordance that could help to discriminate between A0 and AA and the imputation scenario with the highest genotypic concordance that could help to discriminate between B0 and BB. This sometimes resulted in the identification of one single biallelic SNP, in perfect LD with the untyped allele of the triallelic SNP that could be used to discriminate both between A0 and AA and between B0 and BB genotypes.

Appendix B

Consequences of miscalling null-alleles in case-control studies

If the presence of a null-allele is not recognized this will have consequences for case-control association studies. The easiest case is when the null-allele is itself the risk allele. If it is not recognized as such, the SNP will give no signal at all when assuming the A0 and B0 genotypes confer the same risk. However, it is likely these SNPs will be removed from the analysis as HWE deviations are expected to appear and lower call rates will become apparent.

It is more complicated where allele A is the risk allele. Taking the above scenario, we can calculate the odds ratio (OR) of allele A versus non-allele A for the situations where the null-allele is recognized and not recognized. For simplicity, we will limit ourselves to a dominant and a recessive model. In the dominant model, for the observed OR (allele A versus non-allele A) where the null-allele is not recognized, we get:

$$OR_A(obs) = \frac{\gamma\left[(\alpha-1)(p_B+2p_0)+(\alpha\gamma-1)p_A\right]}{(\alpha\gamma-1)(2p_0+\gamma p_A+p_B)}$$

And if the null-allele is typed correctly:

$$OR_A(real) = \frac{\gamma\left[(\alpha-1)(p_B+p_0)+(\alpha\gamma-1)p_A\right]}{(\alpha\gamma-1)(p_0+\gamma p_A+p_B)}$$

where p_A, p_B, p_0 are the allele frequencies of the respective alleles, γ is the disease risk for genotypes not containing A, and $\alpha\gamma$ is the disease risk for individuals carrying one or two A-alleles. Note the difference of $2p_0$ and p_0 in both denominator and numerator between the two equations.

For the recessive model, where penetrance for AA-homozygotes is still $\alpha\gamma$, and penetrance for all other genotypes is α :

$$OR_A(obs) = \frac{(\alpha-1)(p_B+2p_0+\gamma p_A)}{(\alpha-1)(2p_0+p_B)+(\alpha\gamma-1)p_A}$$

And if the null-allele is typed correctly:

$$OR_A(real) = \frac{(\alpha-1)(p_B+p_0+\gamma p_A)}{(\alpha-1)(p_0+p_B)+(\alpha\gamma-1)p_A}$$

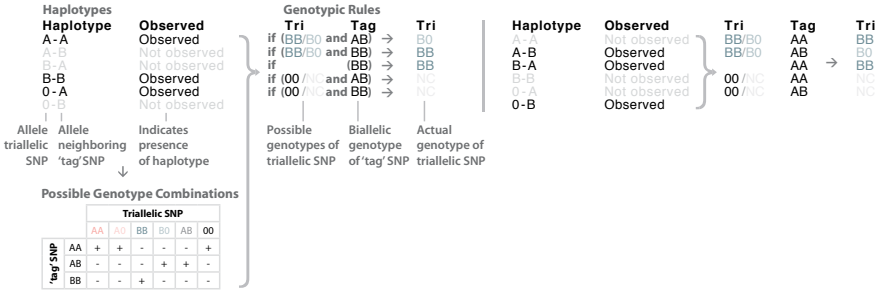
Figure 5 depicts the consequences of mistyping on the observed OR: OR is overestimated for the dominant model, and underestimated for the recessive model. The amount of over- or underestimation depends on the relative penetrance (γ) of the risk allele and the null-allele frequency.

Figure 9

Imputation scenarios

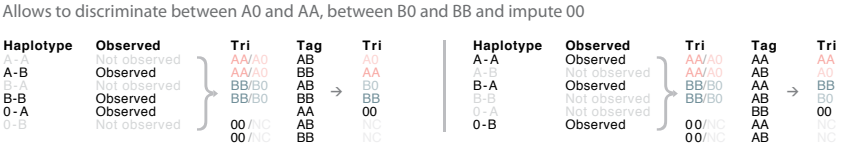
When assuming all alleles have a non-zero frequency for both the triallelic SNP and the neighboring biallelic SNP, and that some LD is present (i.e. at least one haplotype has not been observed), there are 24 different imputation scenarios possible. For 10 of these scenarios the biallelic SNP can help to discriminate between B0 and BB and/or between A0 and AA for the triallelic SNP. For the first imputation scenario a detailed description of this procedure is provided: With this set of observed and unobserved haplotypes, a limited number of genotype combinations exist. This allows for deducing a set of genotypic rules that can help to discriminate between B0 and BB genotypes for the triallelic SNP, based on the genotype of the neighboring biallelic SNP.

Strong LD between untyped allele of triallelic SNP and neighboring biallelic 'tag' SNP
Allows to discriminate between B0 and BB



Strong LD between untyped allele of triallelic SNP and neighboring biallelic 'tag' SNP
Allows to discriminate between A0 and AA

Perfect LD between untyped allele of triallelic SNP and neighboring biallelic 'tag' SNP ($r^2 = 1$)
Allows to discriminate between A0 and AA, between B0 and BB and impute 00



Some LD, but no capability to discriminate between A0 and AA or between B0 and BB

Haplotype	Observed
A-A	Not observed
A-B	Observed
B-A	Not observed
B-B	Observed
0-A	Observed
0-B	Observed

Tri
AA/A0
BB/B0

Tag		Tri
AB	→	A0
AB	→	B0
AA	→	00

Haplotype	Observed
A-A	Not observed
A-B	Observed
B-A	Observed
B-B	Observed
0-A	Not observed
0-B	Observed

Tri
BB/B0
00 /NC
00 /NC

Tag		Tri
AA	→	AA
AA	→	BB
AA	→	NC
AB	→	NC

Haplotype	Observed
A-A	Observed
A-B	Observed
B-A	Observed
B-B	Not observed
0-A	Observed
0-B	Not observed

Tri
AA/A0
00 /NC
00 /NC

Tag		Tri
BB	→	AA
BB	→	AA
AB	→	NC
BB	→	NC

Haplotype	Observed
A-A	Observed
A-B	Observed
B-A	Observed
B-B	Observed
0-A	Not observed
0-B	Observed

Tri
AA/A0
BB/B0
00 /NC
00 /NC

Tag		Tri
AA	→	AA
AA	→	BB
AA	→	NC
AB	→	NC

Haplotype	Observed
A-A	Observed
A-B	Observed
B-A	Not observed
B-B	Observed
0-A	Observed
0-B	Observed

Tri
BB/B0

Tag		Tri
AB	→	B0

Haplotype	Observed
A-A	Observed
A-B	Not observed
B-A	Observed
B-B	Not observed
0-A	Observed
0-B	Observed

Tri
AA/A0
BB/B0

Tag		Tri
AB	→	A0
AB	→	B0
BB	→	00

Haplotype	Observed
A-A	Observed
A-B	Not observed
B-A	Observed
B-B	Observed
0-A	Observed
0-B	Not observed

Tri
BB/B0
00 /NC
00 /NC

Tag		Tri
BB	→	BB
BB	→	BB
AB	→	NC
BB	→	NC

Haplotype	Observed
A-A	Not observed
A-B	Observed
B-A	Observed
B-B	Not observed
0-A	Observed
0-B	Observed

Tri
AA/A0
BB/B0

Tag		Tri
AB	→	A0
AB	→	B0

Haplotype	Observed
A-A	Observed
A-B	Observed
B-A	Not observed
B-B	Observed
0-A	Not observed
0-B	Observed

Tri
AA/A0
00 /NC
00 /NC

Tag		Tri
AA	→	AA
AA	→	AA
AA	→	NC
AB	→	NC

Haplotype	Observed
A-A	Observed
A-B	Observed
B-A	Observed
B-B	Not observed
0-A	Observed
0-B	Observed

Tri
BB/B0

Tag		Tri
AB	→	B0

Haplotype	Observed
A-A	Observed
A-B	Observed
B-A	Observed
B-B	Observed
0-A	Observed
0-B	Not observed

Tri
AA/A0
BB/B0
00 /NC
00 /NC

Tag		Tri
BB	→	AA
BB	→	BB
AB	→	NC
BB	→	NC

Haplotype	Observed
A-A	Observed
A-B	Observed
B-A	Observed
B-B	Not observed
0-A	Observed
0-B	Observed

Tri
AA/A0
BB/B0

Tag		Tri
AB	→	A0
AB	→	B0

Haplotype	Observed
A-A	Observed
A-B	Not observed
B-A	Not observed
B-B	Observed
0-A	Observed
0-B	Observed

Tri
AA/A0
BB/B0
00 /NC
00 /NC

Tag		Tri
BB	→	AA
BB	→	BB
AB	→	NC
BB	→	NC

Haplotype	Observed
A-A	Observed
A-B	Observed
B-A	Observed
B-B	Not observed
0-A	Observed
0-B	Observed

Tri
AA/A0
BB/B0

Tag		Tri
AB	→	A0
AB	→	B0

Haplotype	Observed
A-A	Observed
A-B	Observed
B-A	Observed
B-B	Not observed
0-A	Observed
0-B	Observed

Tri
BB/B0

Tag		Tri
AB	→	B0

Haplotype	Observed
A-A	Observed
A-B	Not observed
B-A	Observed
B-B	Observed
0-A	Observed
0-B	Observed

Tri
AA/A0
BB/B0
00 /NC
00 /NC

Tag		Tri
BB	→	AA
BB	→	BB
AB	→	NC
BB	→	NC

Haplotype	Observed
A-A	Observed
A-B	Not observed
B-A	Not observed
B-B	Observed
0-A	Observed
0-B	Observed

Tri
AA/A0
BB/B0

Tag		Tri
AB	→	A0
AB	→	B0

Haplotype	Observed
A-A	Observed
A-B	Not observed
B-A	Observed
B-B	Observed
0-A	Observed
0-B	Observed

Tri
AA/A0

Tag		Tri
AB	→	A0

Haplotype	Observed
A-A	Not observed
A-B	Observed
B-A	Observed
B-B	Observed
0-A	Observed
0-B	Observed

Tri
AA/A0

Tag		Tri
AB	→	A0

4 Systematic genotype-phenotype analysis of autism susceptibility loci implicates additional symptoms to co-occur with autism

Submitted

Lude H. Franke[1,§], Jacobine E. Buizer-Voskamp[1,2,§], Wouter G. Staal[2], Emma van Daalen[2], Chantal Kemner[2], Roel A. Ophoff[1], Jacob A.S. Vorstman[2], Herman van Engeland[2], Cisca Wijmenga[1,3]

1 Complex Genetics Section, DBG-Department of Medical Genetics, University Medical Centre, Universiteitsweg 100, 3584 CG, Utrecht, The Netherlands

2 Rudolf Magnus Institute of Neurosciences, Department of Child & Adolescent Psychiatry, University Medical Centre, Heidelberglaan 100, 3584 CX, Utrecht, The Netherlands

3 Genetics Department, University Medical Center Groningen, 9700 RB Groningen, The Netherlands.

§ These authors contributed equally to this work

Summary

Many genetic studies in autism have been performed, resulting in the identification of multiple linkage regions and cytogenetic aberrations, but little unequivocal evidence for the involvement of specific genes exists. Enhanced phenotyping of autistic individuals, by identifying novel symptoms in these patients, improves understanding and diagnosis, but also helps to define genetically more homogeneous groups of patients, improving the potential to detect these causative genes. To identify these symptoms, we hypothesized that for some susceptibility loci, autism resembles a contiguous gene syndrome, caused by aberrations within multiple (contiguous) genes, which jointly increase autism susceptibility. This would result in various different clinical manifestations that might be rather atypical, but might also be specific to autism. To test this hypothesis and to identify these symptoms, thirteen susceptibility loci, identified through genetic linkage and cytogenetic analyses, were systematically analyzed. The Online Mendelian Inheritance in Man database was used to identify syndromes caused by mutations in the genes residing in each of these loci. Subsequent analysis of the symptoms expressed within these disorders, allowed us to identify 33 symptoms that were over-represented in previous reports mapping to these loci. Through permutation it was established this number of over-represented symptoms was significantly higher than expected ($P = 0.037$). Some of these symptoms, including seizures and craniofacial abnormalities, support our hypothesis as they are already known to co-occur with autism. These symptoms, together with ones that have not previously been described to co-occur with autism, can be considered for use as in- or exclusion criteria towards defining etiologically more homogeneous groups for molecular genetic studies of autism.

Introduction

Autism spectrum disorder (ASD) is characterized by deviations and delays in the development of reciprocal social interaction and communication, in combination with restricted and repetitive behaviors and interests. These clinical features manifest in the first three years of life (Diagnostic and Statistical Manual of Mental disorders, Fourth edition, Text Revision, DSM-IV-TR, APA, 2000). The prevalence of the broad autism spectrum has recently been estimated to be approximately 1% of the childhood population[1-3], while the prevalence rate for autism is estimated at approximately 4 per 1,000 births[1;4;5]. Phenotypically, autism is very heterogeneous, with varying degrees of severity and associated intellectual functioning[6;7]. The large variety of neuropathologic changes and the variability seen across subjects imply that autism is also etiologically heterogeneous[8;9].

Cumulative evidence from family and twin studies suggests that genetic factors play an important role in the pathology of autism[10-12]. The genetic contribution to autism has been estimated to be as high as 90 percent[7;12-16] Despite the considerable heritability, the mode of transmission is not clear. Findings of cytogenetic abnormalities and single gene disorders associated with autism indicate that the disorder is genetically complex, involving multiple (interacting) loci[7;11;13]. Although no susceptibility loci have been consistently replicated, the overlap in linkage findings from genome scans suggests various regions that could harbor autism susceptibility genes. Loci that have been found in at least two independent linkage studies are in the regions 2q, 3q25-27, 3p25, 6q14-21, 7q31-36 and 17q11-21[7]. However, these loci are rather broad, each containing hundreds of genes, of which multiple genes have been implicated to play a role in autism. Among these, other genes have been identified from independent association studies[12;13;15-17] but no gene has been unequivocally shown to contribute to autism susceptibility.

Recently, evidence has appeared that small cytogenetic aberrations, including duplications, deletions, and copy number variations (CNVs), might play important roles in autism[18;19]. Methods for directly detecting CNVs genomewide provide a powerful alternative to traditional gene-mapping approaches for discovering susceptibility genes in autism[18;20]. Results from a recent CNV analysis suggest that lesions at many different loci can contribute to autism, a result consistent with the findings from cytogenetic studies, as well as consistent with the failure to find causal variants[18].

The results of all molecular genetic studies point to a genetic model of multiple genetic variants that supposedly can interact in various ways with regard to the phenotypic expression of autism[7]. While it can be that within each of these loci one single gene is affected that raises susceptibility to autism, we hypothesized that for some of these loci autism resembles a contiguous gene syndrome, caused by aberrations within multiple (contiguous) genes. As a result, multiple genes are affected in their function, resulting in phenotypes and symptoms that might be specific to autism, but might also be quite atypical.

To substantiate this hypothesis, we sought for evidence that certain atypical symptoms co-occur with autism: For a set of loci that have already been implicated in autism[19] we systematically investigated all positional candidate genes and determined what symptoms are usually caused by aberrations within each of these genes. Subsequently we assessed for each identified symptom whether more than one of the autism susceptibility loci could cause this symptom and determined whether the amount of loci that could cause this symptom (through affected genes within these loci) was significantly higher than expected by chance (Figure 1).

The identification of certain symptoms, reported more often in these loci than expected, would substantiate this hypothesis, and additionally might help to identify symptoms that have not yet been described to co-occur with autism and which could be relevant for the clinic. Additionally, the presence or absence of these symptoms within patients could help to define genetically more homogeneous groups of individuals which are useful for follow-up research[21-23].

Figure 1. Overview of identification of over-represented symptoms in autism loci

For many loci, associated with autism, no genes have been unequivocally shown to be associated. We have assumed that a systematic analysis of symptoms caused by aberrations in positional candidate genes in these loci, might reveal symptoms that are present more often than one would expect by chance, and thus might co-occur with autism. First, loci identified through cytogenetic and linkage analysis are used as input (Step 1). In this example three loci have been identified, each of them causing known diseases or syndromes when mutated (Step 2). For each disorder, we subsequently determine the associated symptoms (Step 3). MeSH is then used for the generation of standardized codes, which are hierarchically organized, allowing to be both specific and generic at the same time (e.g. 'spine' is specific, 'bone and bones' is generic) (Step 4). Once all the symptoms have been recoded, it can be determined what symptoms are expressed per locus (Step 5), allowing for the identification of over-represented ones (e.g. 'bone and bones') (Step 6).

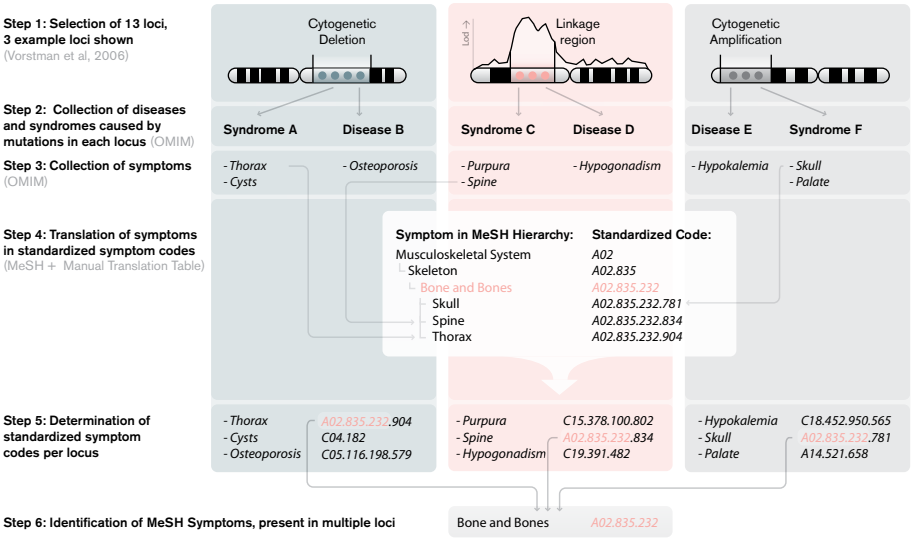


Table 1. Autism susceptibility loci included in the analysis

The chromosomal location is provided for each locus. If a locus has been identified through linkage analysis, the microsatellite markers are given in brackets. The chromosomal location provides base-pair boundaries for each locus. The total numbers of genes for the loci are given, as well as the total number of genes in all the loci and the linkage results/cytogenetic regions of interest.

Locus		Chromosomal Location	No of Genes	Evidence for inclusion
1q42.2	(D1S1656)	217,212,087 - 237,212,087	138	Buxbaum *et al*, 2004[25]
2q31.1	(D2S2188)	165,430,238 - 185,430,238	109	IMGSAC, 2001[26]
2q37		233,875,000 - 243,020,000	77	Vorstman *et al*, 2006[19]
3q26.32	(D3S3037)	168,924,373 - 188,924,373	120	Auranen *et al*, 2002[24]
5p15		0 - 16,900,000	65	Vorstman *et al*, 2006[19]
7q22.1	(D7S477)	90,370,231 - 110,370,231	193	IMGSAC, 2001[26]
7q36.2	(D7S2462)	142,999,403 - 162,999,403	105	Auranen *et al*, 2002[24]
15q11-14		18,940,000 - 31,390,000	65	Vorstman *et al*, 2006[19]
17q11.2	(5-HTTLPR)	15,406,471 - 35,406,471	272	McCauley *et al*, 2004[27]
18q21-23		41,800,000 - 73,160,000	117	Vorstman *et al*, 2006[19]
22q11.2		16,970,000 - 20,830,000	74	Vorstman *et al*, 2006[19]
22q13.3		44,555,000 - 49,550,000	51	Vorstman *et al*, 2006[19]
Xp22		0 - 24,700,000	122	Vorstman *et al*, 2006[19]
		Total number of genes	1,508	

Materials and methods

Definition of susceptibility loci

Loci for autism were selected based on evidence from both linkage and cytogenetic studies. Of all linkage studies that we had analyzed previously[19], four studies that had at least one locus with a multipoint logarithm of the odds (LOD) score (MLS) above 3.0, were included in the analysis[24-27]. Given that no unequivocal method of defining the extent of the region is provided in the literature, boundaries of the linkage regions were pragmatically defined at both ends of a 20 MB base pair block centered around the most significantly linked marker in each locus.

Definition of the Cytogenetic Regions of Interest (CROIs) was based on criteria that have been described before[19]. In short, regions on the human genome where multiple overlapping cytogenetic abnormalities co-occurred with an autism phenotype were identified through an extensive literature search. If a cytogenetic region had been described in over five cases with high quality phenotypes (defined as 3 or more on a five-point scale[19]) and showed no chromosomal mosaicisms, the locus was included for analysis. In order to focus on loci that contained unknown disease genes, cases with an already well-described gene mutation as the most likely genetic cause for autism were excluded as well (for example, patients with fragile-X syndrome caused by *Fmr1* mutations). In total we defined 13 loci, of which six were based on linkage peaks with at least a LOD score above 3.0 and of which seven were based on cytogenetic data (Table 1). The NCBI V35 assembly was used to physically map all markers, probes and banding information.

Identification of syndromes and subsequent symptoms caused by aberrations in loci

The Online Mendelian Inheritance in Man (OMIM) database catalogues the majority of all known diseases that have genetic components, providing extensive information on both clinical aspects and the genetic basis of these syndromes. Based on physical mapping information of mutations in OMIM, we determined which syndromes were caused by aberrations that were (partly) overlapping with each of the 13 loci (Figure 1). OMIM provides a clinical synopsis describing the core symptoms caused by each disorder. As this information is both well organized and extensive, we chose this repository as the basis for collecting symptom information for each syndrome. We included only the core clinical manifestation information, and not the entries contained in the "miscellaneous", "molecular basis", and "inheritance" sections, because these never describe actual symptoms. For each entry only the complete text was used, to prevent subsets of phrases being incorrectly attributed (e.g. "spot quality assessment" was taken as a whole, because "spot" can be interpreted to be a symptom in 'Exanthema'). Subsequently, the Medical Subject Headings (MeSH) vocabulary[28] was used to code these symptoms displayed within disorders in a standardized way. This transformation could be applied as the MeSH ontology is hierarchically organized, allowing one to describe specific symptoms (e.g. 'spine', MeSH code 'A02.835.232.834'), but be generic at the same time ('spine' is part of parent MeSH term 'bone and bones', MeSH code 'A02.835.232.834'). As such, slightly different but related symptoms (e.g. 'skull', 'spine', and 'thorax') all share a more generic parent MeSH term ('bone and bones'), which enabled us to associate these symptoms with each other through a common parent term.

To ensure the automatic assignment of clinical synopsis information to MeSH terms was performed with high accuracy, we also manually assigned all the symptoms for the syndromes contained in the 13 loci to MeSH terms. This manual curation resulted in a conversion table, which maps clinical synopsis entries to known MeSH terms. This allowed for the automatic extraction of information, by text mining[29-32] OMIM and MeSH, and through the conversion table, increased the yield of clinical synopsis assignments to MeSH.

Analysis of over-represented symptoms in loci

We then traversed all MeSH terms, both including those that had been explicitly mentioned, along with their more generic parent

Table 2. Overview of difficulties for text mining in OMIM and utilizing MeSH

Although text mining and natural language processing have gained in attention recently, there are still numerous practical problems to deal with in OMIM and MeSH. Commonly observed difficulties, along with examples, are shown

Difficulty	Description
Difficulties for text mining in OMIM	
Clinical synopsis not designed to be easily machine interpretable	Sometimes, the clinical synopsis contains symptoms such as 'Heart: Prolonged QTc interval; T-wave abnormalities'. Having computers interpret this as something which has to do with MeSH term 'ECG abnormalities', is difficult.
Non-standardized method for describing phenotypes	In some cases, limited clinical synopsis field is present, whereas various symptoms are described in the 'clinical features' part of the full-text OMIM record. Additionally, the clinical synopsis field is not consistent in describing phenotypes. Sometimes different phrasing exists for nearly identical symptoms, such as 'Height: short stature' and 'Height: adult height reduced; final adult height less than 152cm'.
Minor spelling errors within OMIM	In the clinical synopsis sometimes spelling errors are present, such as 'hypereflexia' instead of 'hyperreflexia', 'congential' instead of 'congenital', and 'defeciency' instead of 'deficiency'.
Difficulties with utilizing MeSH	
Symptoms not present in MeSH	Various symptoms are not present in MeSH, such as 'short stature', 'broad nasal bridge' or 'striae'.
Idiosyncrasies in MeSH	'Microcephaly' (C05.660.207.620) is present in MeSH as a member of 'Craniofacial abnormalities' (C05.660.207). However, 'Macrocephaly' is not present in MeSH. The only solution for including this symptom is to assign it to the generic term 'Craniofacial abnormalities'.
Differences in extensiveness of MeSH	MeSH is not equally extensive for all medical subjects: The 'Respiratory tract diseases' (C08) tree contains many highly specific terms, whereas the 'Mental disorders' (F03) tree only contains terms on the level of individual diseases but not that much on specific psychiatric symptoms. Therefore symptoms such as 'Impaired social smile' can only be assigned to the generic term 'Child behavior disorders'.

and grandparent MeSH terms, and determined in how many loci each MeSH term was reported at least once.
Once this was assessed, we determined whether any of these MeSH terms had been described within more loci than expected by performing a 10,000 fold permutation analysis on the data. In each permutation, the 13 loci were shuffled randomly across the genome and the text mining analysis was performed again on these permuted loci: For each MeSH term the number of shuffled loci in which this term had been described was determined and this number was compared to the original number of loci in which this term had been described. Consequently, after these 10,000 permutations, for each MeSH term an empiric p-value could be determined.

In order to identify potentially common symptoms, we only followed up MeSH terms that were present in at least four loci. As our strategy was to determine potentially relevant novel symptoms in autism, we deemed a symptom interesting when its empirically determined p-value was below 0.05.

We assessed whether the number of identified symptoms with an empiric p-value below 0.05 was significantly more than expected by a 1,000 fold permutation analysis. We shuffled the loci randomly across the genome and determined for each permutation how many terms had a p-value below 0.05, using the same filtering as we had applied for the original CROIs. This enabled us to empirically determine whether the amount of nominally significantly identified symptoms was more than expected.

Results

An overview of the 13 selected loci is shown in table 1, along with the evidence for their inclusion (linkage results or cytogenetic region of interest). To ensure that the loci that were identified through cytogenetic analyses were potentially specific to autism and were not commonly deleted or duplicated, we investigated each locus in the Database of Common Genetic Variations[33;34]. None of these loci were known to contain aberrations in healthy individuals as extensive as the ones observed within autism patients.

Once these loci had been defined, OMIM was assessed to determine which known syndromes are caused by mutations in each of these loci. Subsequent analysis of the clinical synopsis information for each syndrome and mapping to MeSH terms using automatic translation provided by MeSH, allowed us to extract a standardized set of symptoms. Assignment of symptoms present within the clinical synopsis for the syndromes in our loci through the use of the manually curated conversion table resulted in the assignment of over 500 extra symptoms to MeSH terms. Although this increase of assignment was considerable and as accurate as possible, mining OMIM and mapping of symptoms to MeSH terms was sometimes problematic, as outlined in table 2.

Once all the syndromes had been processed, we assessed per MeSH term the number of loci in which this term was mentioned. In order to establish whether any term was over-represented, *i.e.* present in more loci than expected by chance, a permutation analysis was performed, which allowed for the determination of an empiric p-value for each term (Figure 1; Table S1).
As we had manually translated the clinical synopsis information for the syndromes which mapped within our 13 loci to MeSH terms, we wanted to ensure that clinical information for syndromes residing outside these loci could also be mapped using this translation table. If a slightly different phrasing of these symptoms had been used in syndromes that we had not manually assessed, as they mapped outside our 13 loci, this could influence the accuracy of the empirically determined p-value. This was, however, not the case, as the results from an analysis that relied entirely on the automatic translation of clinical synopsis symptoms to MeSH terms (Table S1) gave comparable results to an analysis which included the manual assignment of symptoms to MeSH terms (Table S2).

Since autism, Asperger's disorder and RETT syndrome had already been described (OMIM numbers 209850, 608636, 607373, 300495, 312750, 608638, 300497) in four out of the 13 loci, this allowed for an initial validation of our method. Symptoms mentioned in the clinical synop-

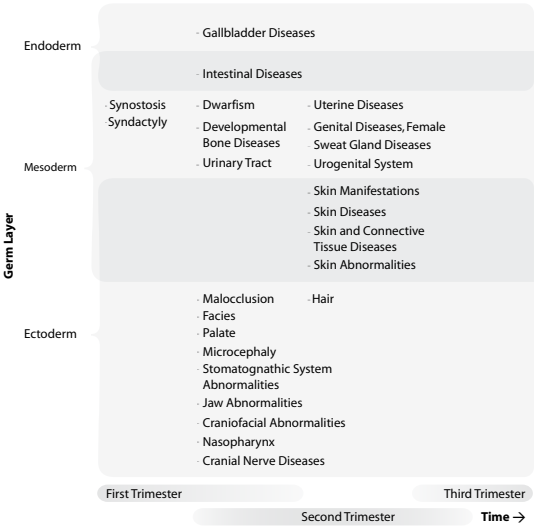
Table 3. Significantly over-represented symptoms mentioned in at least four loci

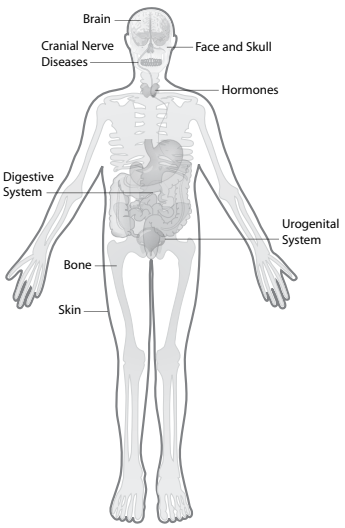
Over-represented symptoms mentioned in at least four loci with an empiric p-value < 0.05 are shown. Symptoms indicated in bold are significantly over-represented and already known to be involved in autism

MeSH number	MeSH description	# Mentioned in different syndromes in 13 CROIs	Empiric P-value
C23.888.885	Skin manifestations	11	0.00369963
C13.371.852	Uterine diseases	5	0.00519948
C07.793.494	Malocclusion	6	0.00629937
A17.360	Hair	12	0.01359864
A14.521.658	Palate	8	0.01409859
C23.550.291.812	Facies	8	0.01479852
C05.660.207.620	Microcephaly	11	0.01579842
C23.300.175	Calculi	4	0.01659834
C07.650	Stomatognathic system abnormalities	13	0.01979802
C10.597.617	Pain	6	0.02319768
C19.391.482	Hypogonadism	6	0.02349765
C17.800	Skin diseases	13	0.02349765
C05.116.099.343	Dwarfism	12	0.02379762
C13.371	Genital diseases, female	10	0.02419758
C17	Skin and connective tissue diseases	13	0.02479752
C05.116.099	Bone diseases, developmental	13	0.02729727
C17.800.946	Sweat gland diseases	4	0.02929707
C05.660.207	Craniofacial abnormalities	13	0.03089691
C05.116.198.579	Osteoporosis	8	0.03109689
A04.623.557	Nasopharynx	6	0.03169683
C18.452.950.565	Hypokalemia	4	0.03189681
C05.500.460	Jaw abnormalities	11	0.03279672
C06.130.564	Gallbladder diseases	4	0.03309669
C10.292	Cranial nerve diseases	13	0.03559644
C10.228.140.490.631	Seizures	13	0.04059594
A05.810	Urinary tract	9	0.04109589
A05	Urogenital system	9	0.04139586
C10.228.140.490	Epilepsy	13	0.04269573
C04.588	Neoplasms by site	8	0.04379562
C05.116.099.370.894	Synostosis	7	0.04479552
C07	Stomatognathic diseases	13	0.04809519
C05.116.099.370.894.819	Syndactyly	6	0.04809519
C06.405.469	Intestinal diseases	9	0.04859514

Figure 2. Overview of over-represented symptoms in autism loci

Over-represented symptoms (empiric p-value < 0.05), present in at least four loci, are shown along with the responsible organs. When possible, symptoms were assigned to a trimester of pregnancy and germ layer. The majority of symptoms are of ectodermal origin, while the majority of affected organs develop in the first to second trimesters





sis information for these syndromes could be attributed to the MeSH term 'Child behavior disorders', for which the empirically determined p-value was 0.01 (Table S2). In order to prevent a bias towards autism syndromes and symptoms already described in OMIM, we excluded these syndromes, along with autism-related syndromes that were defined within OMIM (OMIM numbers 606053, 609378, 611015, 611016, 605309, 608049, 610676, 610836, 300425, 610838, 300496, 610908, 300672, 300624, 608631, 300494, 609954, 608781) – but which mapped outside our 13 loci – from further analyses (Table 3).

Subsequent inspection of the most significantly over-represented symptoms (Table 3) suggests many of these are related (Table 4). Notable are epilepsy/seizures and craniofacial abnormalities, as these have been implicated before in autism[35;36]. Furthermore, the results indicate that most of these symptoms affect tissues that are of ectodermal origin (Figure 2a). They develop in the first and second trimester of pregnancy and affect many organs (Figure 2b).

Discussion

Through text mining of syndromes caused by aberrations in 13 linkage regions and CROIs, this study suggests that various symptoms co-occur with autism which have not yet been widely studied or described before: We found 33 symptoms which were present in these regions more often than expected by chance (nominal empiric p-value < 0.05). However, as many symptoms had been assessed we had to account for multiple testing issues. To do this, we performed an additional analysis to determine whether the number of symptoms that had been found overrepresented (33) was significantly higher than expected. This was indeed the case (empiric p-value = 0.037), indicating that various of the reported symptoms are likely to reflect true positive findings. These observations support our hypothesis that autism might partly be a contiguous gene syndrome, in which the function of multiple positional candidate genes within susceptibility loci is affected. This would result in various different clinical manifestations that might be quite atypical, but jointly might also be able to cause autism like features, which is supported by reports on Xp22.3 deletions, in which patients show the variable association of apparently unrelated clinical manifestations[37;38]. Jointly, the multiple genes with their resulting clinical phenotypes could increase the probability of developing autism. For some of the co-occurring symptoms, evidence already exists that they indeed play a role in autism. The most prominent are epilepsy/seizures and craniofacial abnormalities, which have been mentioned before as possible genetically informative phenotypes in autism[35;36]. Epilepsy is one of the best known and validated associations with autism[39-45]. It is much more common in people with autism than in the general population and, vice versa, it appears that autism and autistic-like conditions are more common in people with epilepsy. Recent studies suggest that more than one-third of the children with autism develop epilepsy[39;43;45]. About 15-20% of all people with autism had seizures before the age of three years[40]. Prevalence rates of epilepsy and the types of seizures seem to depend on the level of mental retardation, age, and incidence of regression[39;42;45]. Not surprisingly, this comorbidity led to researchers proposing that these diseases share common pathophysiological mechanisms[41-43;45]. The observed over-representation of seizures within this study supports these hypotheses because our method assumes that the same genetic background can yield both autism and other symptoms.

Minor physical anomalies, such as craniofacial abnormalities, in association with autism, have also been mentioned frequently[46-55]. Anomalies of the eyes, ears and face (like a broad nasal bridge) have especially been implicated in autism[48;49;52], because of the common ectodermal origin of these anomalies and the brain[47], and the close relationship between cerebral and craniofacial development[48]. Numerous case reports of thalidomide-induced autism suggest abnormal development very early in the gestation, resulting in craniofacial abnormalities[7;46;50;51;53-56]. Although most of these physical anomalies are also sometimes observed in other developmental disorders and in normally developing children

Table 4. Clustering of significantly over-represented symptoms in at least four loci

Relationships between different over-represented symptoms are shown. Symptoms indicated in bold are already known to be involved in autism. Some closely related symptoms were combined (indicated by * and §)

Face & Skull	**Skin diseases**	**Bone diseases**	**Brain & Nerve dysfunction**	**Digestive system diseases**	**Urogenital diseases**	**Metabolic / Endocrine disorders**
Malocclusion Hair Palate Facies Microcephaly Stomatognathic system abnormalities Craniofacial abnormalities Nasopharynx Jaw abnormalities Stomatognathic diseases	Skin manifestations Pain* Skin diseases Skin/connective tissue diseases Sweat gland diseases	Pain* Dwarfism Bone diseases, developmental Osteoporosis Synostosis Syndactyly	Cranial nerve diseases* Seizures Epilepsy	Gallbladder diseases Neoplasms by site§ Intestinal diseases	Uterine diseases Calculi Genital diseases, female Urinary tract Urogenital system	Hypokalemia Hypogonadism
References						
46-55;57	64-66	67;68	51-54;63;39-45	58;59;61;62		

* Original OMIM symptoms: bone pain; back pain; burning of skin

§ Original OMIM symptoms: neoplasms in colon, liver, biliary tract, and gastrointestinal tract

as well[57], craniofacial abnormalities might be potentially interesting[50;52;56], because of their higher frequencies in autistic patients. Many parents report gastrointestinal symptoms in their autistic child[58], which is in line with the digestive system disease symptoms we report. Although gastrointestinal problems are also fairly common in normally developing children, it has been estimated that they affect 46% to 84% of autism patients[59;60]. Chronic diarrhea, increased bile fluid output, constipation, and increased intestinal permeability are the most frequently mentioned abnormalities in autistic children[58-62].

Limited evidence is available for the involvement of cranial nerves in autism. While not convincing, a few studies on thalidomide-induced autism have suggested that the cranial nerves could be dysfunctional[51;53;54;63]. In this form of autism, individuals showed abnormalities in eye movement and facial expression. Other support comes from the observation that the exposure period for thalidomide autistic individuals is during days 20 and 24 gestation. Few neurons form in this period, but the motor neurons of the cranial nerves are a notable exception. Interestingly, these nerves operate the muscles of the ears, jaw, throat, tongue, face and eyes.

For other symptoms, such as skin-, bone-, and urogenital problems, hypokalemia and hypogonadism, there is little evidence for an association with autism. While skin symptoms, such as eczema[64-66], and occasionally bone problems[67;68] have been reported, evidence for the presence of other symptoms is not available.

However, most of these 33 symptoms seem to affect organs that are of ectodermal origin (Table 4). Additionally, for most of these organs, the critical periods of development are during the first and second trimesters of pregnancy, which is in accordance with previous hypotheses about the etiology of autism[46;48;53;63;69]

When taking this supportive evidence into account, many of the symptoms we have found over-represented might indeed play a role in autism and could be considered for follow-up research to determine whether they could function as in- or exclusion criteria for defining etiologically more homogeneous subgroups of autistic patients.

Limitations of our study

While this method has identified various symptoms that are likely to co-occur with autism, we are aware of a number of limitations in our methodology. One important issue is that this study does not unequivocally prove that these symptoms, for which there is no evidence in the literature, are truly associated with autism. It could also be that they have never been studied, since in the clinical setting most attention is usually devoted to a triad of features: social impairments, communication impairments, and restricted repetitive behaviors and interests. Another issue is how to determine what appropriate criteria for including a susceptibility locus are. We tried to do this as careful as possible, but it is possible that some are false-positives. Other loci may well have been overlooked. Apart from these statistical power issues, there are no clear definitions on how to determine the exact boundaries of linkage regions and there is no consensus on whether to include only linkage regions that have shown significant linkage, or to also include loci that were suggestive of linkage. The cytogenetic regions of interest show comparable problems: how many overlapping cases are required in order to consider regions interesting is somewhat arbitrary, and again, how to define the boundaries of these loci accurately is open to discussion.

Not without their own problems are the use of OMIM and MeSH: as OMIM was designed to be interpreted by humans, there was not an immediate need to use a standardized system for coding phenotypes. Consequently, when performing automated text mining in OMIM, various problems became apparent (Table 2): no symptom coding is used, there are sometimes spelling errors in the described symptoms, and very different phrasing is occasionally used to describe the same concept, which makes translating these symptoms into standardized MeSH terms very difficult at times. Another problem is that MeSH differs in extensiveness for the various fields of medicine it covers.

This is particularly detrimental for behavioral symptoms, which lack detailed MeSH equivalents. Although manual curation partly overcame these problems, as it enabled the assignment of a substantial number of extra symptoms to MeSH terms, a more standardized method for describing symptoms both in OMIM and in MeSH is desired. Various authors have already noted that standardization would help considerably in increasing the yield and accuracy of these types of analyses[29;31;70].

Recently, some studies have been published that also use text mining to associate different types of information, a few of which also take OMIM and MeSH into account: Van Driel *et al* (2006)[31] associated different phenotypes with each other using OMIM and MeSH; Butte *et al* (2006)[29] associated phenotypes with expression data, and Lage *et al* (2007)[32] have associated syndromes with protein complexes through text mining of OMIM and protein-protein interaction studies. However, as far as we are aware, no study has utilized OMIM and MeSH to assess whether there are any symptoms over-represented in multiple loci that have been implicated in complex diseases, such as autism, to provide leads for the involvement of unreported symptoms in these disorders.

While much work remains to be done in order to validate the actual co-occurrence of these symptoms in autistic patients, this study might be useful in pointing to ways for better characterizing patients, thereby providing new avenues for genetically informative phenotypes, which could lead to the identification of etiologically more homogeneous groups in patients and increase the statistical power to detect genetic associations. In addition, this method can easily be applied to other psychiatric disorders, as the input for our method consists solely of a set of susceptibility loci and an optional OMIM 'Clinical Synopsis to MeSH term conversion' table. This method will allow researchers to gain insight into the potential involvement of unreported symptoms associated with other psychiatric disorders as well.

Acknowledgments

We thank Jackie Senior, Sasha Zhernakova, Madelien van de Beek, members of the Complex Genetics Section, the Department of Human Genetics, and the Department of Child & Adolescent Psychiatry for critically reading the manuscript. This study was supported by a Fellowship from the Netherlands Genomics Initiative and a grant from the Celiac Disease Consortium, an innovative cluster approved by the Netherlands Genomics Initiative and partially funded by a Dutch government grant (BSIK03009).

References

1 Baird G, Simonoff E, Pickles A, *et al*: Prevalence of disorders of the autism spectrum in a population cohort of children in South Thames: the Special Needs and Autism Project (SNAP). Lancet 2006; 368: 210-5.

2 CDC: Prevalence of autism spectrum disorders - autism and developmental disabilities monitoring network, six sites, United States, 2000. MMWR 2007; 56:(SS-1): 1-11.

3 CDC: Prevalence of autism spectrum disorders - autism and developmental disabilities monitoring network, 14 sites, United States, 2002. MMWR 2007; 56:(SS-1): 12-28.

4 Baird G, Charman T, Baron-Cohen S, Cox A, Swettenham J, Wheelwright S: A screening instrument for autism at 18 months of age: a 6-year follow-up study. J Am Acad Child Adolesc Psychiatry 2000; 39: 694-702.

5 Bertrand J, Mars A, Boyle C, Bove F, Yeargin-Allsopp M, Decoufle P: Prevalence of autism in a United States population: the Brick Township, New Jersey, Investigation. Pediatrics 2001; 108: 1155-61.

6 Persico AM, Bourgeron T: Searching for ways out of the autism maze: genetic, epigenetic and environmental clues. Trends Neurosci 2006; 29: 349-58.

7 Freitag C: The genetics of autistic disorders and its clinical relevance: a review of the literature. Mol Psychiatry 2007; 12: 2-22.

8 Newschaffer CJ, Fallin D, Lee NL: Heritable and nonheritable risk factors for autism spectrum disorders. Epidemiol Rev 2002; 24: 137-53.

9 Herbert M, Russo J, Yang S, *et al*: Autism and environmental genomics. Neurotoxicology 2006; 27: 671-84.

10 Borgatti R, Piccinelli P, Passoni D, *et al*: Relationship between clinical and genetic features in "Inverted Duplicated Chromosome 15" patients. Pediatr Neurol 2001; 24: 111-6.

11 Coon H: Current perspectives on the genetic analysis of autism. Am J Med Genet Part C 2006; 142C: 24-32.

12 Muhle R, Trentacoste SV, Rapin I: The genetics of autism. Pediatrics 2004; 113: 472-86.
13 Bacchelli E, Maestrini E: Autism spectrum disorders: molecular genetic advances. Am J Med Genet Part C2006; 142C: 13-23.
14 Bailey A, Le Couteur A, Gottesman II, *et al*: Autism as a strongly genetic disorder: evidence from a British twin study. Psychol Med 1995; 25: 63-77.
15 Folstein SE, Rosen-Sheidley B: Genetics of autism: complex aetiology for a heterogeneous disorder. Nat Rev Genet 2001; 2: 943-55.
16 Wassink TH, Brzustowicz LM, Bartlett CW, Szatmari P: The search for autism disease genes. Ment Retard Dev Disabil Res Rev 2004; 10: 272-83.
17 Polleux F, Lauder JM: Toward a Developmental Neurobiology of autism. Ment Retard Dev Disabil Res Rev 2004; 10: 303-17.
18 Sebat J, Lakshmi B, Malhotra D, *et al*: Strong association of de novo copy number mutations with autism. Science 2007; 316: 445-9.
19 Vorstman JAS, Staal WG, van Daalen E, van Engeland H, Hochstenbach PFR, Franke L: Identification of novel autism candidate regions through analysis of reported cytogenetic abnormalities associated with autism. Mol Psychiatry 2006; 11: 18-28.
20 Zhao X, Leotta A, Kustanovich V, *et al*: A unified genetic theory for sporadic and inherited autism. Proc Natl Acad Sci USA 2007; 104: 12831-6.
21 Szatmari P, Maziade M, Zwaigenbaum L, *et al*: Informative phenotypes for genetic studies of psychiatric disorders. Am J Med Genet Part B 2007; 144B: 581-8.
22 Bearden CE, Freimer NB: Endophenotypes for psychiatric disorders: ready for primetime? Trends Genet 2006; 22: 306-13.
23 Hasler G, Drevets WC, Gould TD, Gottesman II, Manji HK: Toward constructing an endophenotype strategy for bipolar disorders. Biol Psychiatry 2006; 60: 93-105.
24 Auranen M, Vanhala R, Varilo T, *et al*: A genomewide screen for autism-spectrum disorders: evidence for a major susceptibility locus on chromosome 3q25-27. Am J Hum Genet 2002; 71: 777-90.
25 Buxbaum JD, Silverman J, Keddache M, *et al*: Linkage analysis for autism in a subset families with obsessive-compulsive behaviors: evidence for an autism susceptibility gene on chromosome 1 and further support for susceptibility genes on chromosome 6 and 19. Mol Psychiatry 2004; 9: 144-50.
26 IMGSAC: A genomewide screen for autism: strong evidence for linkage to chromosomes 2q, 7q, and 16p. Am J Hum Genet 2001; 69: 570-81.
27 McCauley JL, Olson LM, Dowd M, *et al*: Linkage and association analysis at the serotonin transporter (SLC6A4) locus in a rigid-compulsive subset of autism. Am J Med Genet Part B 2004; 127B: 104-12.
28 National Institutes of Health. Medical Subject Headings (MeSH(R)). National Library of Medicine 2007Available from: URL: http://www.nlm.nih.gov/mesh
29 Butte AJ, Kohane IS: Creation and implications of a phenome-genome network. Nat Biotechnol 2006; 24: 55-62.
30 Tiffin N, Kelso JF, Powell AR, Pan H, Bajic VB, Hide WA: Integration of text- and data-mining using ontologies successfully selects disease gene candidates. Nucleic Acids Res 2005; 33: 1544-52.
31 van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JAM: A text-mining analysis of the human phenome. Eur J Hum Genet 2006; 14: 535-42.
32 Lage K, Karlberg EO, Storling ZM, *et al*: A human phenome-interactome network of protein complexes implicated in genetic disorder. Nat Biotechnol 2007; 25: 309-16.
33 Redon R, Ishikawa S, Fitch KR, *et al*: Global variation in copy number in the human genome. Nature 2006; 444: 444-54.
34 The centre for applied genomics. Database of Genomic Variants. Department of Genetics and Genomic Biology, MaRS Centre,Canada 2006(Build 36 (Mar 2006)) Available from: URL: http://projects.tcag.ca/variation/
35 Steiner CE: On macrocephaly, epilepsy, autism, specific facial features, and mental retardation. Am J Med Gen Part A 2003; 120A: 564-5.
36 Veenstra-VanderWeele J, Cook EH: Molecular genetics of autism spectrum disorder. Mol Psychiatry 2004; 9: 832.
37 Lonardo F, Parenti G, Luquetti D, *et al*: Contiguous gene syndrome due to an interstitial deletion in Xp22.3 in a boy with ichthyosis, chondrodysplasia punctata, mental retardation and ADHD. Eur J Med Genet 2007; 50: 301-8.
38 Macarov M, Zeigler M, Newman J, *et al*: Deletions of VCX-A and NLGN4: a variable phenotype including normal intellect. J Intellect Disabil Res 2007; 51: 329-33.
39 Canitano R: Epilepsy in autism spectrum disorders. Eur Child Adolesc Psychiatry 2006; 16: 61-6.
40 Danielsson S, Gillberg IC, Billstedt E, Gillberg C, Olsson I: Epilepsy in young adults with autism: a prospective population-based follow-up study of 120 individuals diagnosed in childhood. Epilepsia 2005; 46: 918-23.

41 Gabis L, Pomeroy J, Andriola MR: Autism and epilepsy: cause, consequence, comorbidity, or coincidence? Epilepsy Behav 2005; 7: 652-6.
42 Gillberg C, Coleman M. The biology of the autistic syndromes. 3rd ed. London: Mac Keith Press; 2000.
43 Rossi PG, Posar A, Parmeggiani A: Epilepsy in adolescents and young adults with autistic disorder. Brain Dev 2000; 22: 102-6.
44 Steffenburg S, Steffenburg U, Gillberg C: Autism spectrum disorders in children with active epilepsy and learning disability: comorbidity, pre- and perinatal background, and seizure characteristics. Devel Med Child Neurol 2003; 45: 724-30.
45 Tuchman R, Rapin I: Epilepsy in autism. Lancet Neurol 2002; 1: 352-8.
46 Arndt TL, Stodgell CJ, Rodier PM: The teratology of autism. Int J Dev Neurosci 2005; 23: 189-99.
47 Cantor-Graae E, McNeil TF, Torrey EF, *et al*: Link between pregnancy complications and minor physical anomalies in monozygotic twins discordant for schizophrenia. Am J Psychiatry 1994; 151: 1188-93.
48 Hardan A, Keshavan MS, Sreedhar S, Vemulapalli M, Minshew NJ: An MRI study of minor physical anomalies in autism. J Autism Dev Disord 2006; 36: 607-11.
49 Lauritsen MB, Mors O, Mortensen PB, Ewald H: Medical disorders among inpatients with autism in Denmark according to ICD-8: a nationwide register-based study. J Autism Dev Disord 2002; 32: 115-9.
50 Miles JH, Hillman RE: Value of a clinical morphology examination in autism. Am J Med Genet 2000; 91: 245-53.
51 Rodier PM, Ingram JL, Tisdale B, Nelson S, Romano J: Embryological origin for autism: developmental anomalies of the cranial nerve motor nuclei. J Comp Neurol 1996; 370: 247-61.
52 Rodier PM, Bryson SE, Welch JP: Minor malformations and physical measurements in autism: data from Novo Scotia. Teratology 1997; 55: 319-25.
53 Rodier PM: The early origins of autism. Sci Am 2000; 2: 56-63.
54 Rodier PM: 2003 Warkany lecture: autism as a birth defect. Clin Mol Teratology 2004; 70: 1-6.
55 Wier ML, Yoshida CK, Odouli R, Grether JK, Croen LA: Congenital anomalies associated with autism spectrum disorders. Dev Med Child Neurol 2006; 48: 500-7.
56 Hultman CM, Sparén P, Cnattingius S: Perinatal risk factors for infantile autism. Epidemiology 2002; 13: 417-23.
57 Merks JHM, Özgen HM, Cluitmans TLM, *et al*: Normal values for morphological abnormalities in school children. Am J Med Genet Part A 2006; 140A: 2091-109.
58 Horvath K, Papadimitriou JC, Rabsztyn A: Gastrointestinal abnormalities in children with autistic disorder. J Pediatr 1999; 135: 559-63.
59 Kuddo T, Nelson KB: How common are gastrointestinal disorders in children with autism? Curr Opin Pediatr 2003; 15: 339-43.
60 Erickson CA, Stigler KA, Corkins MR, Posey DJ, Fitzgerald JF, McDougle CJ: Gastrointestinal factors in autistic disorder: a critical review. J Autism Dev Disord 2005; 35: 713-27.
61 Horvath K, Perman JA: Autistic disorder and gastrointestinal disease. Curr Opin Pediatr 2002; 14: 583-7.
62 White JF: Intestinal pathophysiology in autism. Exp Biol Med 2003; 228: 639-49.
63 Miller MT, Strömland K, Ventura L, Johansson M, Bandim JM, Gillberg C: Autism associated with conditions characterized by developmental errors in early embryogenesis: a mini review. Int J Dev Neurosci 2005; 23: 201-19.
64 Gurney JG, McPheeters ML, Davis MM: Parental report of health conditions and health care use among children with and without autism. Arch Pediatr Adolesc Med 2006; 160: 825-30.
65 Titomanlio L, Marzano MG, Rossi E, *et al*: Case of Myhre syndrome with autism and peculiar skin histological findings. Am J Med Genet 2001; 103: 163-5.
66 Whiteley P: Developmental, behavioral and somatic factors in pervasive developmental disorders: preliminary analysis. Child Care Health Dev 2004; 30: 5-11.
67 Bolton P, Powell JE, Rutter M, *et al*: Autism, mental retardation, multiple exostoses and short stature in a female with 46,X,t(X;8)(p22.13;q22.1). Psychiatr Genet 1995; 5: 51-5.
68 Lohiya GS, Tan-Figueroa L, Iannucci A: Identification of low bone mass in a developmental center: finger bone mineral density measurement in 562 residents. Am J Med Dir Assoc 2004; 5: 371-6.
69 Celani G: Comorbidity between autistic syndrome and biological pathologies: which implications for the understanding of the etiology? J Dev Phys Disabil 2003; 15: 141-54.
70 Biesecker LG: Mapping phenotypes to language: a proposal to organize and standardize the clinical descriptions of malformations. Clin Genet 2005; 68: 320-6.

Supplementary table 1

Analysis of over-represented symptoms mentioned in at least four loci, solely relying upon mappings as provided by MeSH, without excluding known autism and Asperger syndromes

MeSH number	MeSH description	# Mentioned in different syndromes in 13 CROIs	Empiric P-value
C17.800	Skin diseases	13	0.00529947
C05.660.207	Craniofacial abnormalities	13	0.00789921
A17.360	Hair	12	0.01009899
C05.660	Musculoskeletal abnormalities	13	0.01049895
C23.888.885	Skin manifestations	10	0.01329867
C05.116.198.579	Osteoporosis	8	0.01589841
C05.660.207.620	Microcephaly	11	0.01609839
C13.371.852	Uterine diseases	4	0.02129787
C05.660.585	Limb deformities, congenital	7	0.02459754
C18.452.950.565	Hypokalemia	4	0.02679732
C17.800.946	Sweat gland diseases	4	0.03019698
A04.623.557	Nasopharynx	6	0.03259674
A05.810	Urinary tract	9	0.04169583
A05	Urogenital system	9	0.04219578
C10.597.617	Pain	4	0.04319568
C05.116.099.370	Dysostosis	10	0.04439556
C05.116.099	Bone diseases, developmental	11	0.04459554
C05.116.198	Bone diseases, metabolic	8	0.04629537
C16.131	Abnormalities	13	0.04759524

Supplementary table 2

Analysis of over-represented symptoms mentioned in at least four loci, relying both upon mappings provided by MeSH and manual curation, without excluding known autism and Asperger syndromes

MeSH number	MeSH description	# Mentioned in different syndromes in 13 CROIs	Empiric P-value
C13.371.852	Uterine diseases	5	0.00369963
C23.888.885	Skin manifestations	11	0.00379962
C07.793.494	Malocclusion	6	0.00649935
F03.550.300	Child behavior disorders	4	0.01009899
A17.360	Hair	12	0.01169883
C23.550.291.812	Facies	8	0.01239876
C05.660.207.620	Microcephaly	11	0.01569843
A14.521.658	Palate	8	0.01579842
C23.300.175	Calculi	4	0.01769823
C07.650	Stomatognathic system abnormalities	13	0.01989801
C17.800	Skin diseases	13	0.0209979
C17	Skin and connective tissue diseases	13	0.02279772
C19.391.482	Hypogonadism	6	0.02439756
C10.597.617	Pain	6	0.02449755
C10.597.606.150.500.550	Language development disorders	7	0.02489751
C13.371	Genital diseases, female	10	0.02679732
C05.116.099.343	Dwarfism	12	0.02779722
C18.452.950.565	Hypokalemia	4	0.02909709
C05.116.099	Bone diseases, developmental	13	0.02979702
C17.800.946	Sweat gland diseases	4	0.0309969
C05.116.198.579	Osteoporosis	8	0.03159684
C06.130.564	Gallbladder diseases	4	0.03249675
C05.500.460	Jaw abnormalities	11	0.03259674
C05.660.207	Craniofacial abnormalities	13	0.03267673
C10.292	Cranial nerve diseases	13	0.03479652
A04.623.557	Nasopharynx	6	0.03559644
C10.228.140.490.631	Seizures	13	0.04149585
C04.588	Neoplasms by site	8	0.04209579
C06.405.469	Intestinal diseases	9	0.04439556
C10.228.140.490	Epilepsy	13	0.04469553
A05.810	Urinary tract	9	0.04489551
A05	Urogenital system	9	0.04509549
C05.116.099.370.894	Synostosis	7	0.04649535
C16.131.831	Skin abnormalities	8	0.04679532
C15.378.100.802	Purpura	4	0.0479952
C17.800.329	Hair diseases	8	0.04889511
C06.405.469.158	Colonic diseases	6	0.04939506

5 Complex nature of genetic variation on gene expression in primary human leucocytes

In preparation

Graham Heap[1], Gosia Trynka[2,§], Ritsert Jansen[3,§], Marcel Bruinenberg[2], Lotte C. Dinesen[4], Karen Hunt[1], Cisca Wijmenga[2,5], David A. van Heel[1], Lude H. Franke[1,2,5]

1 Institute of Cell and Molecular Science, Barts and The London School of Medicine and Dentistry, London, E1 2AT, UK

2 Genetics Department, University Medical Centre Groningen and University of Groningen, 9700 RB Groningen, the Netherlands

3 Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Kerklaan 30, NL-9751 NN Haren, the Netherlands.

4 Gastroenterology Unit, University of Oxford, Oxford OX3 7BN, UK.

5 Complex Genetics Section, DBG-Department of Medical Genetics, University Medical Centre Utrecht, 3584 CG Utrecht, the Netherlands

§ These authors contributed equally to this work.

Summary

Genetic variation influences gene expression and splicing, and can be mapped by genetical genomics approaches. With a view to identifying function of disease associated risk variants arising from genome wide association studies, we correlated gene expression and genetic variation in untouched primary leucocytes from individuals with a quiescent immune mediated disease (celiac disease). We compared our observations in primary leucocytes with EBV-transformed B cell line observations, and increased power to detect non-tissue specific effects in a meta-analysis of these datasets. Peripheral blood RNA was collected using the PAXgene system from 110 unrelated treated UK celiac disease individuals. Samples were analyzed for 257,013 autosomal SNPs (Illumina HumanHap300 BeadChip) and 19,867 transcript levels (Illumina HumanRef-8 v2 BeadChip). In peripheral blood, 2,178 SNP variants influenced gene expression at 658 different transcripts within 500kp of the SNP (*cis* eQTL) at a false discovery rate threshold of 0.05. 204 of these cis eQTLs were also observed in a published dataset of EBV-transformed B cell line RNA in 90 Caucasian HapMap samples, all having an effect in the same allelic direction. Gene expression differences between EBV transformed cell lines and PAXgene primary leucocyte samples predominantly explain the limited overlap in observed *cis*-eQTLs. An overrepresentation of 'defense and immunity' genes was observed in the PAXgene primary leucocyte samples, compared to the EBV transformed cell lines. We identified 16 genes where a SNP significantly affected the expression of multiple probes that mapped to different exons, but for which the allelic directions were opposite, We additionally observed that co-expression for the 658 different transcripts with other genes was generally more easily detectable if the genotypic *cis*-effect was removed. Combined, these findings suggest that several *cis*-eQTLs reflect alternative splice-isoforms rather than overall gene expression level differences. Genetical genomics is a useful tool to enhance functional understanding of genetic variants. Here, the use of primary leucocytes from treated disease individuals may have enriched for celiac disease associated alleles/mutations at known risk loci. Risk variants in two celiac disease associated genes, *IL18RAP* and *CCR3*, both exhibited significant *cis* genotype-expression correlations in this dataset but not in the EBV transformed cell line dataset. More *cis*-eQTLs could be identified in a meta-analysis of datasets of different cell types. At several loci, complex effects of genotype on gene expression were observed.

Introduction

Human gene expression levels have a strong heritable component[1-10]. At some genes, the variance in gene expression levels is up to 11 fold greater between unrelated individuals, than between identical twins[2].

Quantitative mRNA levels are key regulators of phenotype and represent a link between genetic variation and phenotypic alterations. A term first introduced by Jansen & Nap[11], *genetical genomics* aims to identify the genetic variants that affect gene expression. By treating gene expression as a quantitative trait it is possible to correlate gene transcript expression levels with genomic locations such that expression quantitative trait loci (eQTLs) can be identified[11; 12]. In the human genome, *cis* associations, where a genetic variant affects a transcript that maps to the same locus, have been predominantly reported[7-10]. *Trans* effects, where the genetic variant is distant to the transcript loci, are much harder to convincingly identify due to inherent multiple testing problems. Analysis of *trans* effects involves several magnitudes more statistical tests than for *cis* effects. Although individual studies have reported human *trans* associations, no effects have been convincingly replicated in multiple studies identified for the same transcript and variant[7-10].

Quantitative transcript expression and genotype relationships can be investigated via linkage or association based methodologies. Linkage studies use a genome wide series of markers in recombinant in-bred lines or families, to follow the heritability of a trait, whilst association studies usually compare large number of single nucleotide polymorphisms (SNPs) to transcript levels from unrelated individuals. Despite its extensive use in yeast, plant, mouse and rat models[13-17], transcript expression and genotype correlation studies have only recently been performed in humans[13].

Human association studies have centered on RNA obtained from leukocytes, predominantly Epstein-Barr virus (EBV) transformed B cell lines from HapMap individuals[7;9]. Large-scale genetical genomics linkage study have recently identified more than 1000 *cis* regulatory loci across the genome, in primary cryopreserved human leukocytes[8; 10]. However, the former studies are limited by their ability to assess transcripts that are significantly expressed at the point of RNA isolation. Cell extraction methods, cryopreservation and EBV transformation all affect individual mRNA expression levels. These variations in RNA analysis make the choice of analysis tissue of paramount importance. Linkage studies are limited by identification of broad genomic regions rather than variants in more precisely defined LD blocks.

Here we present the first association based genetical genomics study using primary cell RNA from peripheral blood in 110 celiac disease patients. Immediate RNA preservation during blood sampling (using the PAXgene system) represents nearly *in vivo* human physiological gene expression. Celiac disease is a common (1% prevalence), inflammatory condition of the small intestine induced by intake of gluten in wheat, rye and barley. A strong genetic component has been established for disease with a monozygotic concordance of 75%[18] and 90% of cases possessing the HLA haplotype HLA-DQ2.5[19] and the remainder mostly have HLA-DQ8[20]. Despite the role of the HLA, the risk of disease is still greater in HLA matched monozygotic twins compared to HLA matched dizygotic twins[21]. We recently performed a genome wide association and replication study using single nucleotide polymorphisms (SNPs) and identified an additional eight susceptibility loci that predispose to celiac disease[22;23]. Likely candidate immune system genes (expressed in leukocytes) from seven of these loci suggested the feasibility of genetical genomics approaches using peripheral blood to assess the biological function of celiac disease associated risk variants.

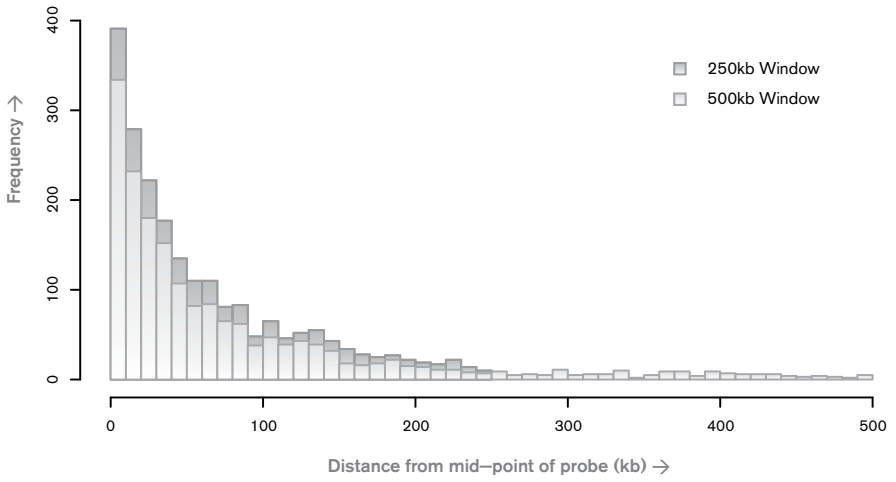
We show that numerous *cis*-eQTLs can be identified through an expression analysis of peripheral blood RNA. We also show many of these are only detectable in peripheral blood RNA, and not in EBV-transformed B cell lines. Through a meta-analysis of these two datasets we identified numerous additional *cis*-eQTLs. We show that for some *cis*-eQTLs genetic variation does not lead to

Table 1: Overview of detected cis-eQTLs in two different populations

Population	110 celiac disease samples (PAXgene peripheral blood)		90 Caucasian HapMap samples (EBV-transformed B cell lines)	
Expression data	Illumina HumanRef-8 v2 Whole Genome BeadChip (19,867 mapped transcripts)		Illumina HumanRef-6 v1 Whole Genome BeadChip (44,791 mapped transcripts)	
Genotype data	Illumina Infinium HumanHap300 BeadChip (313,505 SNPs)		Subset of all HapMap Genotypes present on Illumina Infinium HumanHap300 BeadChip (313,505 SNPs)	
	FDR 0.01	FDR 0.05	FDR 0.01	FDR 0.05
Spearman's Correlation	$p < 6.56 \times 10^{-6}$	$p < 5.44 \times 10^{-5}$	$p < 4.18 \times 10^{-6}$	$p < 3.46 \times 10^{-5}$
Number of performed tests	1,850,599	1,850,599	3,820,148	3,820,148
Number of detected *cis*-eQTLs	1127	1823	1398	2134
Number of unique probes	340	576	432	678
Number of unique genes	326	554	348	547
Number of unique SNPs	1067	1728	1262	1929

Figure 1

Cumulative genomic distance distribution between SNP and probe midpoint for significant cis-eQTLs (FDR = 0.01).



overall gene expression changes, but rather leads to shifts in the types of different splice isoforms that are produced.

Materials & Methods

Study population

In this study we enrolled treated celiac disease individuals, leading to an enrichment of celiac disease associated genetic variants. Celiac disease is an immune mediated disease, dominated by T_H1 cytokine response. Given the importance of tissue specific RNA profiles, we felt that peripheral blood was an appropriate medium to study to investigate celiac disease associated genetic variants.

Peripheral blood from gluten-free diet treated celiac disease patients was extracted at the St Bartholomew's Hospital, London and the John Radcliffe Hospital, Oxford. All patients were long-term celiac disease patients with a small bowel endoscopic biopsy diagnosis of celiac disease. Enrolled patients had a median age of 51, a median age at diagnosis of 42, a male to female sex ratio of 1:3, and a median length of treatment on a gluten free diet of 9.4 years. Patients responding to a gluten free diet show no detectable inflammation, and their peripheral blood samples are essentially immunologically indistinguishable from healthy controls. In total 115 patients and 22 healthy controls were enrolled in this study. Blood was collected with fully informed consent and permission from the medical ethics committees from both hospitals.

PAXgene RNA Extraction

2.5 ml of peripheral blood was collected into a PAXgene tube (Becton Dickinson, UK, 762165). PAXgene vials were chosen to prevent density gradient centrifugation, immortalization or *in vitro* cell culture that can lead to artifacts in RNA profiles.

PAXgene tubes were mixed gently and incubated at room temperature for two hours. After collection, tubes were frozen at -20°C for at least 24 hours followed by storage at -80°C. The process of RNA extraction was followed according to the protocol of the manufacturer. All reagents were obtained from the PAXgene Blood RNA isolation kit (Qiagen, UK, 762174). RNA was quantified using the Nanodrop (Nanodrop Technologies, USA). Total RNA integrity was analyzed using an Agilent Bioanalyzer (Agilent Technologies, USA) according to manufacturers' instructions.

Anti-sense RNA synthesis, amplification, purification and hybridization

Anti-sense RNA was synthesized, amplified and purified using the Ambion Illumina TotalPrep Amplification Kit (Ambion, USA) following the manufacturers' protocol. Complementary RNA was hybridized to Illumina HumanRef-8 v2 Whole Genome BeadChips and scanned on the Illumina BeadArray Reader. Data was handled through the Illumina BeadStudio Gene Expression module v3.2.

Quality control

Five celiac disease samples were excluded from subsequent analysis due to poor median probe intensity correlation with all other samples or incorrect sex assignment, based on an analysis of all the probes that mapped to the non-pseudoautosomal region of chromosome Y, leaving 110 celiac disease patients for analysis.

Normalization

Expression probes were mapped to the cDNA sequence from Ensembl v45_36g[24] and the NCBI build 36 genome assembly if necessary. Probes that had less than 96% sequence homology or that mapped to multiple loci were removed. Subsequent analyses were confined to autosomal probes, in order to prevent sex specific effects on gene expression. After removal of probes that map to sex chromosomes, data was quantile-quantile normalized[25].

Celiac disease sample genotypes

All celiac disease patients were genotyped as previously described[22] using Illumina Infinium HumanHap300 BeadChips.

Peripheral blood eQTL association analysis and false discovery rate control:

257,013 autosomal SNPs were tested for association with expression levels in the 110 celiac disease samples that met analysis criteria of minor allele frequency (MAF) >

Table 2: Over- and underrepresented biological processes and functions.

Listed are significantly over- and underrepresented biological processes and functions, derived through the Panther Classification System (Binomial test, Bonferroni corrected)

110 celiac disease samples (peripheral blood RNA):

Biological Process	Nr Genes	P-Value	Genes
Immunity and defense	41	2.8×10^{-6}	ABCC3, ADORA3, AHSA2, C3AR1, C4BPA, CARD15, CAT, CCR3, CD9, CFD, CLEC2B, CLEC4C, CLEC4F, EPHX2, F2RL1, FCRL5, GPX7, GRB2, GSTM3, GSTM4, GSTT1, HP, IL18RAP, IRF5, KIR3DL1, LGALS2, LRRC8C, LYZ, MARCO, NFE2L3, NUP88, OAS2, ORM1, OSIL, PF4V1, PHCA, PPIE, PPIL3, PRDX5, RAD51C, SLC11A1
Mesoderm development	19	1.2×10^{-2}	ACP5, BTN3A2, BTN3A3, BTNL3, CLEC3B, COL18A1, FBN2, FHL3, HIP1, MYO18B, MYOM2, NFE2L3, NKX3-1, PF4V1, PVALB, RPS6KA2, STAT6, TPM2, ZHX2
Amino acid metabolism	10	2.3×10^{-2}	ADI1, AMDHD1, ASNSD1, CBS, CRAT, FAS, GRHPR, PAPSS1, PHGDH, SHMT1
Muscle contraction	9	3.1×10^{-2}	C3AR1, MYO18B, MYOM2, NEBL, PVALB, SRI, TAGLN, TNNT1, TPM2
Detoxification	6	4.4×10^{-2}	ABCC3, GSTM4, GSTM3, GSTT1, EPHX2, GPX7
Amino acid biosynthesis	5	4.6×10^{-2}	ASNSD1, CBS, GRHPR, PAPSS1, PHGDH
Molecular Function	**Nr Genes**	**P-Value**	**Genes**
Hydrolase	31	1.6×10^{-7}	ADARB2, AMDHD1, BST1, CHI3L2, DPYSL4, EPHX2, EXOSC6, EXOSC9, FAS, FAS, GLB1L, GNB5, HYAL3, INPP5E, LIPA, LYPLAL1, LYZ, MAN1A1, MAN2C1, NT5C3, NT5C3L, NUDT2, PADI2, PDE6H, PHCA, PLA2G4C, PTER, RAD51C, RNASE2, RNASE3, VNN1, VNN3
Transferase	26	1.7×10^{-3}	AYTL1, CAT, CHPT1, CHST13, CRAT, EXOSC6, FAS, GCAT, GCNT2, GSTM3, GSTM4, GSTT1, HYAL3, LYCAT, MAP3K2, OAS2, PAPSS1, PASK, POLR2J, RPS6KA2, SHMT1, SPDY1, TPST1, TRMT12, UAP1L1, WBSCR27

90 HapMap CEU samples (EBV transformed B cell line RNA):

Biological Process	Nr Genes	P-Value	Genes
T-cell mediated immunity	13	3.5×10^{-3}	CTSS, HLA-B, HLA-C, HLA-DOB, HLA-DQA1, HLA-DQA2, HLA-DRB5, HLA-H, LTBR, LTC4S, SLFN5, SQSTM1, TAPBPL
Coenzyme metabolism	7	1.1×10^{-2}	C9orf95, ECHDC3, HIBCH, MTHFSD, NAPRT1, OXCT2, PDHX
Carbohydrate metabolism	22	1.1×10^{-2}	AIM1, APIP, ATHL1, BLK, BPGM, CHI3L2, CHIA, CRYZ, ECHDC3, EDEM1, EXTL2, FUT10, GAA, HIAT1, HIATL1, HIBCH, IREB2, LDHC, MAN1A2, OXCT2, PDHX, PPP1R3B
Lipid, fatty acid and steroid metabolism	25	3.2×10^{-2}	ACOX3, AMACR, ANXA5, ARSA, ASCC1, CAT, CAV2, CHPT1, ECHDC3, FDX1, HABP4, HIBCH, HSD17B12, IVD, LTC4S, OXCT2, PGS1, PIP5K1C, PIP5K2A, PLA2G4C, PLCB2, PLTP, SLC27A5, SLC37A1, TAP2
Molecular Function	**Nr Genes**	**P-Value**	**Genes**
Major histocompatibility complex antigen	8	1.9×10^{-4}	HLA-B, HLA-C, HLA-DOB, HLA-DQA1, HLA-DQA2, HLA-DRB5, HLA-H, OCIAD2
Hydrolase	29	4.5×10^{-4}	AMDHD1, ARSA, ATHL1, ATP13A1, CHI3L2, CHIA, DCTD, DDX58, DEADC1, DNASE1L3, DPYSL4, ENDOG, FAHD1, GAA, GUF1, MAN1A2, MANEAL, MCMDC1, MTHFD1L, NT5C3L, NUDT2, PLA2G4C, PLCB2, PPA2, PTER, QRSL1, RAD18, RAD51, ZRANB1
Transferase	31	2.0×10^{-3}	CAT, CAT, CCDC126, CDKN1A, CHPT1, ECHDC3, EXTL2, FTSJ2, FTSJ3, FUT10, GSTT1, HIBCH, LCMT1, LTC4S, MAK10, MAP3K2, MGMT, NAPRT1, NMNAT3, NSUN4, OXCT2, PGS1, PIGF, POFUT2, POLR2E, POLR2J, QRSL1, SETD1A, SHMT1, STK25, TGM5, WBSCR27
G-protein coupled receptor	0	1.3×10^{-2}	(Depletion)
Epimerase / racemase	6	2.3×10^{-2}	APIP, AMACR, ECHDC3, ENOSF1, GSTT1, HIBCH

0.1, exact Hardy-Weinberg equilibrium P-Value > 0.0001 and call-rate > 0.95. Analyses were confined to those probe-SNP pairs for which the distance from probe genomic midpoint to SNP genomic location was less than 250kb or 500kb, depending on the analysis performed. To prevent spurious associations due to outliers, a non-parametric Spearman's rank correlation analysis was performed. In order to correct for multiple testing we controlled the false discovery rate (FDR)[26]. Through permutation the Spearman's rank correlation P-value threshold could be determined that corresponded to an FDR of 0.01 and 0.05.

Validation panel: HapMap CEU samples

We compared the identified *cis*-eQTLs in the celiac peripheral blood dataset to a published human genetical genomics dataset[25]. We reanalyzed expression data from EBV-transformed B cell lines (further described as HapMap B cell line dataset) for 90 CEU HapMap samples[3]. Analyses were performed as described for the celiac disease samples. To enable a comparison between the celiac peripheral blood dataset and the HapMap B cell line dataset, only SNPs were tested that had been successfully called within HapMap and that were present on the Illumina HumanHap300 platform. Although this is only a subset of all the SNPs that have been called for these HapMap samples, this subset of SNPs is known to capture most genetic Caucasian variation well[27;28].

Analysis of over- and underrepresented biological processes and function

We investigated over- or underrepresentation of certain biological processes or functions through an analysis of all significant *cis*-eQTL genes using the Panther Classification System[29] (Binomial P-Value, Bonferroni corrected).

Co-expression analysis

As many *cis*-eQTLs have been detected, but only a few trans-eQTLs have been found[7-10], we performed a co-expression analysis to gain insight into the background of this phenomenon. We investigated whether the identified *cis*-eQTLs (probe distance of 250kb, FDR of 0.05) showed co-expression with other probes for both the celiac peripheral blood dataset and the HapMap B cell line dataset. We correlated the measured intensity levels with all other probes that mapped on different chromosomes, through a Spearman's rank correlation analysis. For each *cis*-eQTL the 100 highest observed absolute correlation coefficients were recorded. Subsequently we removed the genotypic effect on the measured probe intensity for each of the *cis*-eQTLs. For each *cis*-eQTL probe we rank transformed the measured probe intensity, determined the slope and changed the intensities for the three genotype groups, such that no longer a correlation between the genotype and measured probe intensity was present. We then assessed co-expression between these 'decorrelated' expression intensities, and all other probes that mapped to different chromosomes and recorded the 100 highest absolute correlation coefficients for each *cis*-eQTL. We used a two-sided Wilcoxon-Mann-Whitney test to assess whether co-expression for *cis*-eQTLs with other probes generally improved after the *cis*-effect on the measured intensity levels was removed.

Results

Gene expression in celiac disease versus healthy control samples

To obtain the most accurate reflection of mRNA levels in peripheral blood leukocytes, whole blood RNA was immediately fixed during venepuncture in PAXgene vials, giving a reflection of *in vivo* RNA expression from whole blood. One hundred and fifteen treated celiac patients, all of whom were successfully treated and compliant with a gluten free diet for at least six months, were enrolled along with 22 healthy control samples. Seventy percent of celiac disease patients were female, (66.6% of adult cases diagnosed with celiac disease in the population are female[19]). No known inflammatory disease associated cytokines, including *IFNG*[30], *IL18*[31], and *IL2*[32] showed significantly increased expression in celiac versus control samples, as was expected since these patients had been treated with a gluten free diet.

Figure 2

Meta-analysis of identical probes between coeliac peripheral blood and HapMap B cell datasets

Summary of meta-analysis of 4,681 identical probes between the celiac peripheral blood and HapMap B cell line datasets. (at an FDR = 0.05)

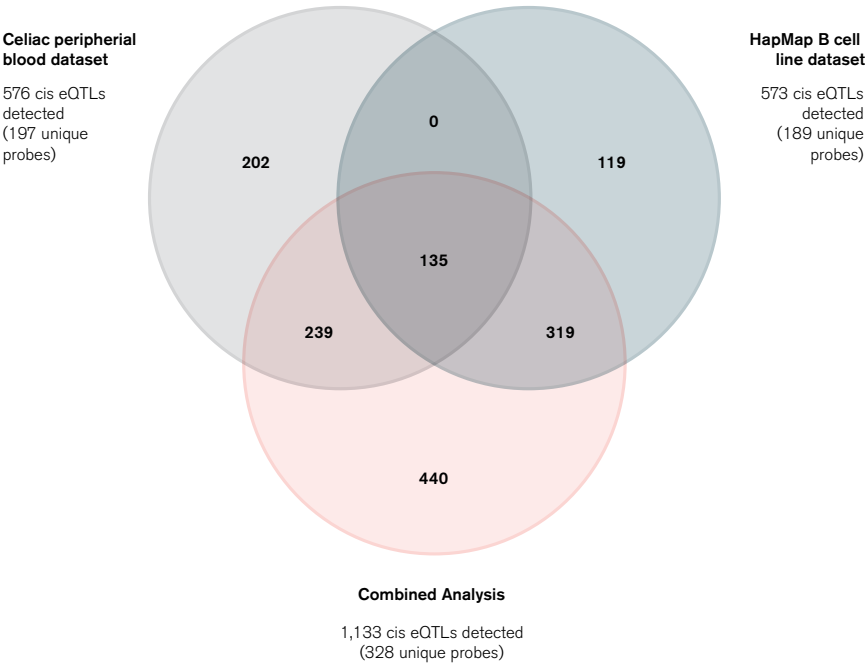
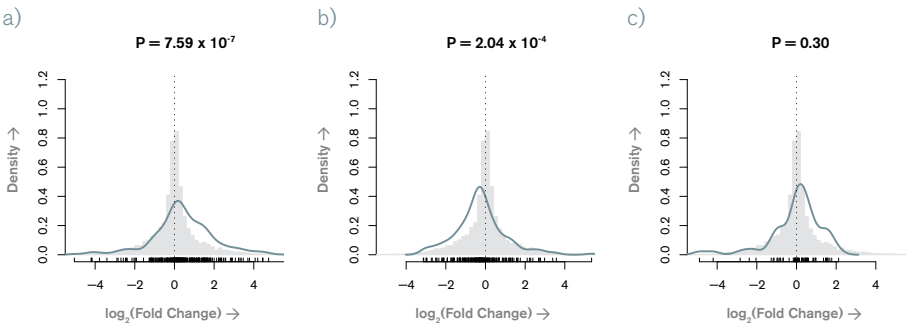


Figure 3

Differential Gene Expression between tissue types results in differential cis-eQTL detection

Differential gene expression between the coeliac dataset and the HapMap dataset is represented as a histogram of log fold. **a)** *cis*-eQTLs detected in coeliac dataset but not in HapMap samples (FDR = 0.01, 500kb window). **b)** *cis*-eQTLs detected in HapMap dataset but not in the coeliac dataset, **c)** *cis*-eQTLs detected in both datasets. P values derived from a Wilcoxon Signed-Ranks Test.



Cis-associations of gene expression with SNPs

For 110 celiac disease samples that passed quality control, both expression and genotype data was available. Gene expression levels for 19,867 transcripts were analyzed for significant genotypic effects at SNPs, mapping within 500kb of the centre of the transcript probes, resulting in 1,850,599 tests. By using a Spearman's rank correlation coefficient statistic and an FDR of 0.01 or 0.05, 1,127 and 1,823 *cis*-eQTLs were detected, respectively (see Table 1).

An identical analysis was performed on publicly available EBV-transformed B cell line expression and genotype data for 90 CEU HapMap individuals[7]. Using this dataset and a SNP-probe distance of 500 kb and FDR of 0.01 or 0.05, 1,398 and 2,134 *cis*-eQTLs were detected, respectively (see Table 1).

Distance between SNPs and probes that constitute the *cis*-eQTLs:

Our analysis was confined to probes that had a midpoint distance to the tested SNPs less than 500kb. Analysis of the significant *cis*-eQTLs SNP-probe distances (Figure 1) suggests few *cis*-eQTLs have been missed by imposing this threshold, as in both datasets for 95% of the *cis*-eQTLs, the SNPs map within 250kb of the probes. As such it is expected that an increase in statistical power can be achieved by reducing the distance to 250kb as less tests will be performed. Indeed, an additional 292 *cis*-eQTLs were identifiable in the celiac peripheral blood dataset while an additional 24 were identified in the HapMap B cell line dataset (controlling the FDR at 0.05 for each analysis) (Figure 1).

Primer Polymorphisms

While it can be assumed that for most of the detected *cis*-eQTLs indeed the probe expression is affected by genetic variation, it can also be that SNPs, mapping to regions to which the probe hybridizes, may affect hybridization efficacies and result in *cis*-eQTLs[33] that are not due to expression differences. For the celiac peripheral blood dataset, 10% of all Illumina HumanRef-8 v2 probes map to regions that contain known dbSNP polymorphisms. For the HapMap B cell line dataset, 20.5% of all Illumina HumanRef-6 v1 probes map to a known SNPs. For the probes that make up the identified *cis*-eQTLs (distance 250kb, FDR = 0.01) this percentage is significantly higher for both the celiac peripheral blood analysis (Fisher's Exact test P=0.0099) and even more pronounced for the HapMap B cell line dataset ($P=1.15 \times 10^{-15}$).

Overrepresented biological pathways

The genes, comprising the significant *cis*-eQTLs, showed an overrepresentation of hydrolase and transferase functions for both datasets, but 'immunity and defense' *cis*-eQTLs genes were more predominantly detected in the celiac disease peripheral blood dataset than in the HapMap B cell line dataset (see Table 2).

Meta-analysis celiac peripheral blood and HapMap B cell lines datasets

4,681 Illumina expression probes had oligonucleotide sequences that are shared between the two different oligonucleotide arrays that had been used. By limiting the analysis to a window size of 250kb and only to SNPs that had been successfully genotyped in both studies, 576 *cis*-eQTLs in the celiac peripheral blood data at a FDR = 0.05 (339, FDR=0.01) were detected. In the HapMap B cell line data, 573 *cis*-eQTLs could be identified at an FDR = 0.05 (339, FDR = 0.01). A combined meta-analysis of both cohorts (Weighted-Z Method) identified 1,133 *cis*-eQTLs at an FDR = 0.05 (428, FDR = 0.01) (see Figure 2). 440 of these were not detected when either dataset was analyzed separately. 135 unique, identical SNP-Probe *cis*-eQTL effects had been identified in both the 110 celiac disease samples and in the 90 B cell line HapMap samples (FDR = 0.05). The combined meta-P-Value for each of these shared *cis*-eQTLs was significant which means that these *cis*-eQTLs all have the same allelic direction.

Differential gene expression between tissue sample types influences *cis*-eQTL detection

Differential gene expression analysis, limited to 12,401 transcripts that map to the

Table 3

14 Genes, containing multiple essayed probes that are affected by SNPs that also affect other probes in the same gene, but with opposite allelic directions

Meta-analysis of different significant but opposite allelic effects of SNPs (FDR = 0.05, 250 kb distance) in the celiac pheripheral blood dataset and HapMap B cell line dataset. Shown are 14 genes that contain probes for either of the two platform but that show opposite allelic directions.

Source	SNP	HUGO	Spearman	Probe	Probe Sequence
Celiac	rs1131383	POLR2J	-0.48	GI_62422568	
HapMap	rs1131383	POLR2J	0.57	GI_21704275	
Celiac	rs1901198	IRF5	-0.60	GI_38683857	
HapMap	rs1901198	IRF5	0.50	GI_38683858	
Celiac	rs6565724	LOC400566	0.67	GI_62177143	
HapMap	rs6565724	LOC400566	-0.72	GI_37544593	
HapMap	rs6565724	LOC400566	-0.74	GI_42661283	
Celiac	rs2863095	MRPL43	0.45	GI_28872731	
Celiac	rs2863095	MRPL43	-0.49	GI_28872733	
HapMap	rs2863095	MRPL43	0.58	GI_28872731	
HapMap	rs2863095	MRPL43	-0.61	GI_28872733	
HapMap	rs4768933	DIP2B	0.42	GI_17457388	
HapMap	rs4768933	DIP2B	-0.44	GI_39930390	
Celiac	rs10774679	OAS1	-0.54	GI_74229010	
Celiac	rs10774679	OAS1	0.51	GI_74229012	
HapMap	rs3177979	OAS1	0.63	GI_8051620	
HapMap	rs3177979	OAS1	-0.72	GI_8051622	
Celiac	rs1040404	TIPRL	-0.47	GI_73088904	
Celiac	rs1040404	TIPRL	0.50	GI_73088933	
HapMap	rs222851	C17orf81	-0.72	GI_44662825	
HapMap	rs222851	C17orf81	0.64	GI_44662829	
Celiac	rs34374	PAM	0.43	GI_21070979	GGCTACAGTCGAAAAGGGTTTGACCGGCTTAGCA
HapMap	rs34374	PAM	-0.41	GI_21070979	GCCAGTGTCTTTCTTTGGTGCCTTTCCTGTTCAGCA
HapMap	rs2838859	POFUT2	-0.48	GI_34147486	
HapMap	rs2838859	POFUT2	0.45	Hs.300736	
Celiac	rs7084722	PTER	-0.43	GI_47933342	
HapMap	rs7084722	PTER	0.48	GI_20070185	
Celiac	rs10503170	MYOM2	0.43	GI_4505314	ATTTTCACGGGTGTGGGCACATGGGTGTGGCACCT
HapMap	rs10503170	MYOM2	-0.42	GI_4505314	TTTACACGAGGGTAGACGGCAGATGCCTGACAGA
Celiac	rs11680305	ADI1	-0.42	GI_8922761	CCGGTGGTGTGATGATGCCATATACCGCAGGGCTT
HapMap	rs11680305	ADI1	0.41	GI_8922761	GAGCTCCCACCCTAAGGGGCACACACTGAGTTGC
Celiac	rs2395185	HLA-DRB5	-0.41	GI_26665892	GGCTCTTATTCTTCCACAAGAGAGGACTTTCTCAG
HapMap	rs2395185	HLA-DRB5	0.47	GI_26665892	ACGGCCTCCCATGCATCTGTACTCCCCCTGTGTGC

same exons, between the celiac peripheral blood and the HapMap B cell line dataset, reveals, as expected, marked gene expression differences. *cis*-eQTLs that have only been significantly detected in the celiac peripheral blood dataset, reflect probes that generally show higher expression in celiac peripheral blood samples compared to HapMap B cell line samples (Wilcoxon Signed-Ranks test, $P=7.6 \times 10^{-7}$) (see Figure 3). Conversely, probes that comprised HapMap B cell line specific *cis*-eQTLs were higher expressed than within the celiac peripheral blood dataset ($P=2.0 \times 10^{-4}$). As expected, probes that comprised *cis*-eQTLs that were common to both data sets did not show differences in expression ($P = 0.30$). Noteworthy is that *IL18RAP* and *CCR3*, both recently found to be associated with celiac disease[23], exhibited significant eQTLs in the celiac peripheral blood dataset but not in the HapMap B cell line dataset. *cis*-effects for these two genes could be detected as these two genes showed markedly upregulated expression in the celiac peripheral blood dataset (log 4.6 and 3.8 fold increase, respectively), which indicates that the analysis of different cell type is valuable to gain insight in the potential functional consequences of disease associated genetic variants.

Significant opposite allelic directions for probes, mapping in the same genes.

The observed concordance in allelic direction for the 135 significant *cis*-eQTLs, detected in both the 110 celiac peripheral blood and the 90 HapMap B cell line samples was perfect when probe sequences were identical.

We expanded our analysis to all probes that were present on either of the two oligonucleotide arrays that we analyzed. This enabled us to identify 14 genes (see Table 3), to which probes with different sequences map that comprise significant *cis*-eQTLs but that show opposite allelic directions.

Detection of *trans*-eQTLs and biological pathways using genetical genomics

One goal of genetical genomics is to uncover previously unknown biological pathways. Genetic variation, affecting the expression of a gene in *trans*, suggests that a biological relationship exists between the two loci. To assess this, considerable amounts of effort have been devoted to the development and application of statistical frameworks that are capable of detecting these *trans*-eQTLs. Detection of *trans*-eQTLs in human populations has proven less successful than in mouse, rat and plant recombinant inbred lines. It has been suggested that the extensive genetic and environmental diversity between human individuals masks many of the existing *trans*-effects.

We used the celiac disease dataset to assess whether significant *trans*-eQTLs could be identified (Spearman's rank correlation test P-Value $< 1 \times 10^{-10}$, FDR = 0.05). Two trans-eQTLs were detected: rs2318331 (mapping within *COL22A1*) with GI_48093066 (*NBPF3*) and rs12634559 (mapping downstream of *IL1RAP*) with GI_16306505 (*CASP8AP2*). Although IL1-RAP and *CASP8AP2* are interestingly both involved in apoptosis, these *trans*-eQTLs were not detected in the HapMap B cell line dataset, even after relaxing the nominal Spearman rank correlation P-value to 0.05.

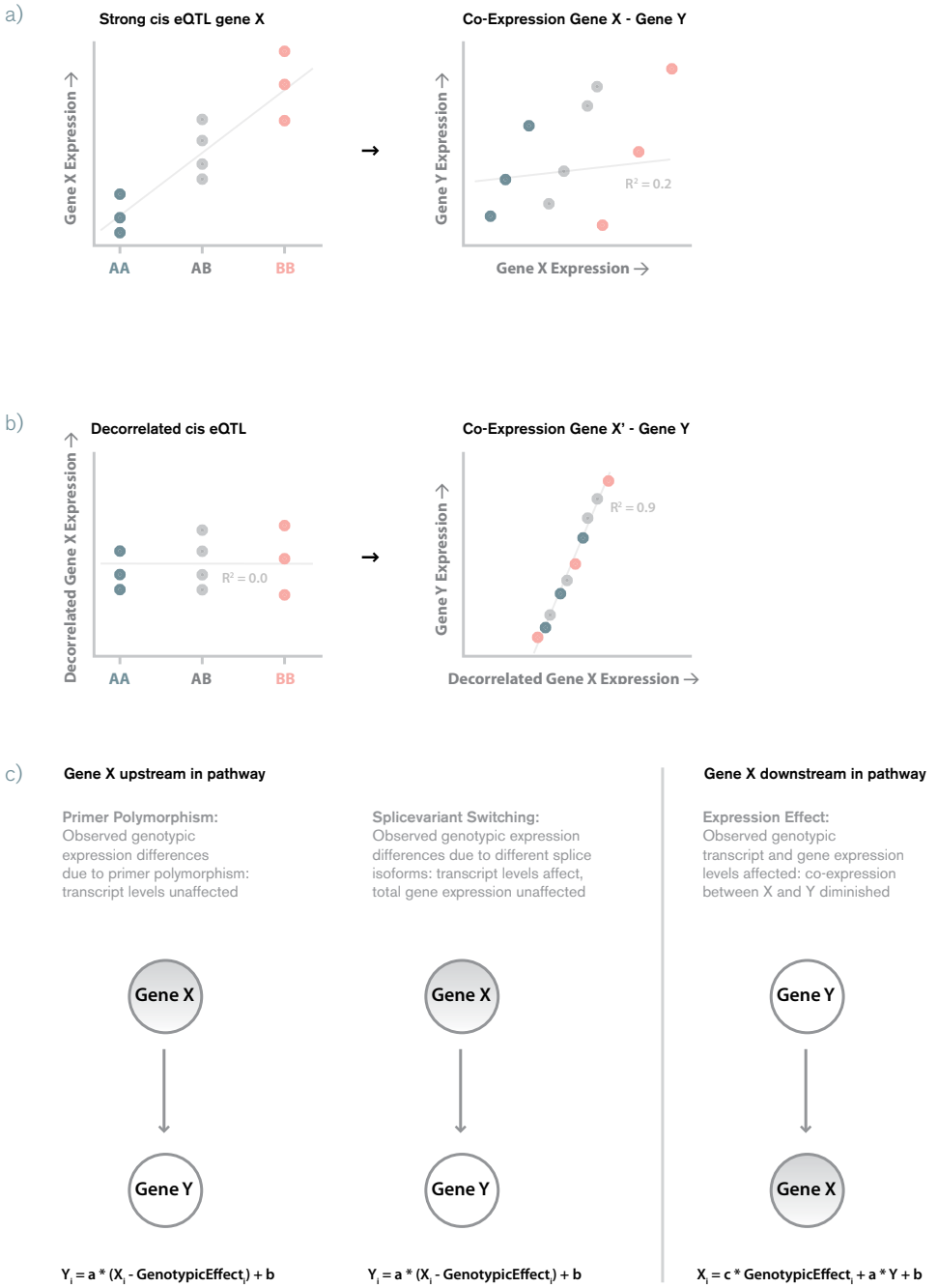
We had observed that a considerable proportion of transcripts were under genetic *cis*-control. Some of these *cis*-eQTLs are due to probe polymorphisms while others are due to differential splicing. If the resulting splice isoforms perform the same function with the same efficacy or are both non-functional, such genotypic variation is not likely to have functional effects on other genes, and *trans*-effects will consequently be not be detected. In the observed data this hypothesis would mean that for some of the genes the measured probe intensity does not reflect the actual overall gene expression intensity but rather isoforms of alternate splice events. We hypothesized that the genetic variation for the strong *cis*-eQTLs in some cases does not really influence overall gene expression. To investigate this we removed the genotypic effect on the probe intensities for each of the significant *cis*-eQTLs, such that the correlation between the corresponding *cis*-SNP and the probe intensities disappeared (Figure 4).

Probe dbSNP Polymorphisms	
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
CTGAGGGCAGTGACCA	0
ATTCTTAGCCTGTGG	0
	1
	0
	0
	1
GGACGTGTGCAGCAT	1
GAGTGGGTTGGCAGAC	1
GCTTCTGTCAAGTGT	1
TTATGCCACTTCCTTG	0
GCCCTGGTTGCTACC	3
CACAAATGCACTACG	8

Figure 4

***cis*-effects obscure detection of co-expression with other genes**

a) Co-expression for significant cis-eQTLs was determined, resulting in the identification of co-expression pairs with generally low absolute correlation coefficients **b)** Through removal of the genotypic effect, for the large majority of identified cis-eQTLs co-expression is more significantly abundant. **c)** Three scenario's that explain this observation are given.



We then calculated co-expression between the original probe intensity measurements and all other probes (Figure 4A) and between the "decorrelated" probe intensity measurements (for which the genotypic effect had been removed) and all other probes (Figure 4B).

In the celiac peripheral blood dataset for 73% of the identified *cis*-eQTLs (SNP-probe midpoint distance 250kb, FDR = 0.05), co-expression improved. Comparable results were obtained in the HapMap B cell line dataset: For 79% of the detected *cis*-eQTLs co-expression improved. These results were especially pronounced for strong *cis*-eQTLs: In the celiac peripheral blood dataset for 90 out of the 100 (90%) strong *cis*-eQTLs (Spearman's correlation coefficient P-Value $< 1 \times 10^{-9}$) stronger co-expression was observed. For the HapMap B cell line dataset co-expression improved for 75 out of the 81 (93%) strong *cis*-eQTLs (Spearman's correlation coefficient P-Value $< 1 \times 10^{-9}$).

Moreover, the proportion of 80,350 known biological interactions (derived on 17 April 2007 from KEGG, BioGrid, Reactome, BIND, HPRD and IntAct) among the top 100 co-expressed genes for each of these decorrelated *cis*-eQTLs probes is significantly higher than among the top 100 co-expressed genes, based on the original probe intensities (Fisher's exact test P-Value = 3.8×10^{-3} for the celiac peripheral blood dataset and P-Value = 1.35×10^{-4} for the HapMap B cell line dataset).

Discussion

We have demonstrated the use of peripheral blood RNA samples for the detection of *cis*-eQTLs. We have shown that there is strong allelic concordance with *cis*-eQTLs that also had been detected in HapMap EBV-transformed B cell line RNA samples. These results indicate that a meta-analysis with larger sample size and hence statistical power results in a considerable increase in the detected *cis*-eQTL. Additionally, it can be concluded that RNA obtained from different cell types, give *cis*-eQTLs that have consistent allelic directions.

Most of the detected *cis*-eQTLs in these datasets however were only detected in one of the two tissues, suggesting that more insight can be gained in the functional consequences of genetic variation by conducting genetical genomics studies using different types of cells and tissues. The greater number of *cis*-eQTLs detected in the EBV validation set likely represents the increased power from the homogeneous cell type, again underlying the importance of individual cell RNA profiles and regulation upon *cis*-eQTL detection.

It is attractive to assume most of the observed *cis*-eQTLs reflect overall gene expression level alterations. However, we did observe 14 genes (Table 3) where different probes showed significant opposite allelic effects. For at five out of the 14 genes (*POFUT2*, *PTER*, *MYOM2*, *ADI1* en *HLA-DRB5*) polymorphisms within the probe are known to exist within dbSNP. As such for these genes, it can be that for each SNP that affects the two probes in opposite directions, one *cis*-eQTL for instance reflects a real expression difference, whereas the other reflects a hybridization effect.

However, recently Kwan *et al*[34] showed for three of the 14 genes (*IRF5, MRPL43* and *PTER*) using Affymetrix GeneChip Human Exon 1.0 ST Array comparable characteristics. They assumed different exonic effects[26] through an independent validation using quantitative RT-PCR (out of a total of 25 validated genes) and estimated that only 39% of the detected *cis*-eQTLs influence overall gene expression levels. For the remaining *cis*-eQTLs genetic variation results in preliminary terminated transcripts (18%), not initiated transcripts (11%), transcripts that are spliced differentially (26%) or a combination of these (6%).

These estimates are further supported by our observation that for most of the transcripts, constituting the identified *cis*-eQTLs, only limited co-expression with other transcripts was observed. When we removed the genotypic effects, many of these 'decorrelated' transcripts suddenly showed strong co-expression with biologically plausible genes.

One explanation, following our findings and those of Kwan *et al*, could be that differential splicing due to genetic variation might not have functional consequences, although expression differences for individual probes are observed (Figure 4C). Another explanation is that these transcripts map to genes that reside downstream of regulatory genes (Figure 4c). In these occasions it is expected that the large genotypic effects convolute the subtler co-expression that is caused by upstream genes. While another potential explanation is that for some of the *cis*-eQTLs genetic variation has no effect on the expression of downstream genes it can be assumed that for a some of the other identified *cis*-eQTLs primer polymorphisms have affected probe hybridization, and as such actually no expression differences are present (Figure 4c).

These results also explain to some extend why there has been such a discrepancy between the amount of detected human *cis*-eQTLs and trans-eQTLs[7-10], compared to genetical genomics studies in recombinant inbred lines[14; 15]. Although these lines have lost a considerable amount of genetic variation which leads to higher statistical power to detect eQTLs (especially *trans*-eQTLs), it can also be assumed primer polymorphisms are less likely to become apparent than in outbred populations. As it has been one of the goals of genetical genomics to identify biological relationships we suggest that the co-expression analysis we carried out here, might help to uncover these: We have proposed that more known biological relationships can be identified when using genetical genomics to perform conditioned co-expression analyses.

This study shows that PAXgene isolated peripheral blood RNA is a powerful resource for investigating functional consequences of genetic variation. We have shown that for some of the *cis*-eQTLs the functional consequences are more complex than previously assumed. Additionally, these findings imply that biological relationships can be extracted in outbred populations, although in a somewhat different manner than what is commonly used to detect biological relationships through trans-eQTLs in inbred model organisms.

As this study has yet only combined genetics with genomics, we envision more extensive integrative approaches, incorporating *e.g.* epigenetics and proteomics, will help to improve the detection of previously unknown biological pathways.

Acknowledgements

We thank UK clinicians who collected samples, and sample donors. We thank Erik Sluiters for mapping all probe identifiers. Some statistical analyses were performed using the Genetic Cluster Computer in Amsterdam (http://www.geneticcluster.org), which is financially supported by the Netherlands Organization, for Scientific Organization (NWO, grant 480-05-003). We acknowledge funding from Celiac UK; the Netherlands Organization for Scientific Research (NWO, VICI grant 918-66-620); the Celiac Disease Consortium (an innovative cluster approved by the Netherlands Genomics Initiative and partly funded by the Dutch government (grant BSIK03009)); the Netherlands Genomics Initiative (grant 050-72-425 and fellowship grant to L.F.); the Wellcome Trust (GR068094MA Clinician Scientist Fellowship to D.A.v.H.)

References:

1 Cheung VG, Spielman RS (2002) The genetics of variation in gene expression. Nature genetics 32 Suppl:522-525

2 Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, Morley M, Spielman RS (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. Nature genetics 33:422-425

3 Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P, Edwards S, Phillips JW, Sachs A, Schadt EE (2004) Genetic inheritance of gene expression in human cell lines. American journal of human genetics 75:1094-1105

4 Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG (2004) Genetic analysis of genome-wide variation in human gene expression. Nature 430:743-747

5 Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT (2005) Mapping determinants of human gene expression by regional and genome-wide association. Nature 437:1365-1369

6 Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, Akey JM (2007) Gene-expression variation within and among human populations. American journal of human genetics 80:502-509

7 Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavare S, Deloukas P, Hurles ME, Dermitzakis ET (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science (New York, NY 315:848-853

8 Goring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole SA, Jowett JB, Abraham LJ, Rainwater DL, Comuzzie AG, Mahaney MC, Almasy L, MacCluer JW, Kissebah AH, Collier GR, Moses EK, Blangero J (2007) Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nature genetics 39:1208-1216

9 Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M, Lathrop GM, Abecasis GR, Cookson WO (2007) A genome-wide association study of global gene expression. Nature genetics 39:1202-1207

10 Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, et al. (2008) Genetics of gene expression and its effect on disease. Nature 452:423-428

11 Jansen RC, Nap JP (2001) Genetical genomics: the added value from segregation. Trends Genet 17:388-391

12 Alberts R, Fu J, Swertz MA, Lubbers LA, Albers CJ, Jansen RC (2005) Combining microarrays and genetic analysis. Briefings in bioinformatics 6:135-145

13 Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, Friend SH (2003) Genetics of gene expression surveyed in maize, mouse and man. Nature 422:297-302

14 Keurentjes JJ, Fu J, Terpstra IR, Garcia JM, van den Ackerveken G, Snoek LB, Peeters AJ, Vreugdenhil D, Koornneef M, Jansen RC (2007) Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. Proceedings of the National Academy of Sciences of the United States of America 104:1708-1713

15 Keurentjes JJ, Fu J, de Vos CH, Lommen A, Hall RD, Bino RJ, van der Plas LH, Jansen RC, Vreugdenhil D, Koornneef M (2006) The genetics of plant metabolism. Nature genetics 38:842-849

16 Petretto E, Mangion J, Pravanec M, Hubner N, Aitman TJ (2006) Integrated gene expression profiling and linkage analysis in the rat. Mamm Genome 17:480-489

17 Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P, Cookson WO, Taylor MS, Rawlins JN, Mott R, Flint J (2006) Genome-wide genetic association of complex traits in heterogeneous stock mice. Nature genetics 38:879-887

18 Greco L, Romino R, Coto I, Di Cosmo N, Percopo S, Maglio M, Paparo F, Gasperi V, Limongelli MG, Cotichini R, D'Agate C, Tinto N, Sacchetti L, Tosi R, Stazi MA (2002) The first large population based twin study of coeliac disease. Gut 50:624-628

19 van Heel DA, Hunt K, Greco L, Wijmenga C (2005) Genetics in coeliac disease. Best practice & research 19:323-339

20 Karell K, Louka AS, Moodie SJ, Ascher H, Clot F, Greco L, Ciclitira PJ, Sollid LM, Partanen J (2003) HLA types in celiac disease patients not carrying the DQA1*05-DQB1*02 (DQ2) heterodimer: results from the European Genetics Cluster on Celiac Disease. Human immunology 64:469-477

21 Nistico L, Fagnani C, Coto I, Percopo S, Cotichini R, Limongelli MG, Paparo F, D'Alfonso S, Giordano M, Sferlazzas C, Magazzu G, Momigliano-Richiardi P, Greco L, Stazi MA (2006) Concordance, disease progression, and heritability of coeliac disease in Italian twins. Gut 55:803-808

22 van Heel DA, Franke L, Hunt KA, Gwilliam R, Zhernakova A, Inouye M, Wapenaar MC, et al. (2007) A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nature genetics 39:827-829

23 Hunt KA, Zhernakova A, Turner G, Heap GA, Franke L, Bruinenberg M, Romanos J, et al. (2008) Newly identified genetic risk variants for celiac disease related to the immune response. Nature genetics

24 Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, et al. (2007) Ensembl 2007. Nucleic acids research 35:D610-617

25 Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics (Oxford, England) 19:185-193

26 Hochberg Y, Benjamini Y (1990) More powerful procedures for multiple significance testing. Statistics in medicine 9:811-818

27 Pe'er I, de Bakker PI, Maller J, Yelensky R, Altshuler D, Daly MJ (2006) Evaluating and improving power in whole-genome association studies using fixed marker sets. Nature genetics 38:663-667

28 Barrett JC, Cardon LR (2006) Evaluating coverage of genome-wide association studies. Nature genetics 38:659-662

29 Cho RJ, Campbell MJ (2000) Transcription, genomes, function. Trends Genet 16:409-415

30 Anderson RP, van Heel DA, Tye-Din JA, Barnardo M, Salio M, Jewell DP, Hill AV (2005) T cells in peripheral blood after gluten challenge in coeliac disease. Gut 54:1217-1223

31 Lettesjo H, Hansson T, Bergqvist A, Gronlund J, Dannaeus A (2005) Enhanced interleukin-18 levels in the peripheral blood of children with coeliac disease. Clinical and experimental immunology 139:138-143

32 Ciclitira PJ, Ellis HJ (1998) In vivo gluten ingestion in coeliac disease. Digestive diseases (Basel, Switzerland) 16:337-340

33 Alberts R, Terpstra P, Li Y, Breitling R, Nap JP, Jansen RC (2007) Sequence polymorphisms cause many false cis eQTLs. PLoS ONE 2:e622

34 Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, Beaulieu P, Hudson TJ, Sladek R, Majewski J (2008) Genome-wide analysis of transcript isoform variation in humans. Nature genetics

6 A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21

Nature Genetics, 2007 Jul;39(7):827-9.

David A. van Heel[1], Lude H. Franke[2,§], Karen A. Hunt[1,§], Rhian Gwilliam[3,§], Alexandra Zhernakova[2], Mike Inouye[3], Martin C. Wapenaar[4], Martin C. N. M. Barnardo[5], Graeme Bethel[3], Geoffrey K. T. Holmes[6], Con Feighery[7], Derek Jewell[8], Dermot Kelleher[7], Parveen Kumar[1], Simon Travis[9],Julian R.F. Walters[10], David S. Sanders[11], Peter Howdle[12], Jill Swift[13], Raymond J. Playford[1], William M. McLaren[3], M. Luisa Mearin[14,15], Chris J. Mulder[16], Ross McManus[7], Ralph McGinnis[3], Lon R. Cardon[8], Panos Deloukas[3], Cisca Wijmenga[2,4]

1 Centre for Gastroenterology, Institute of Cell and Molecular Science, Queen Mary University of London, London E1 2AT, UK.
2 Complex Genetics Section, Department of Biomedical Genetics, University Medical Center Utrecht, 3584 CG Utrecht, The Netherlands.
3 Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK.
4 Genetics Department, University Medical Center Groningen, 9700 RB Groningen, The Netherlands.
5 Transplant Immunology, Oxford Transplant Centre, Churchill Hospital, Oxford OX3 7LJ, UK.
6 Department of Gastroenterology, Derbyshire Royal Infirmary, London Road, Derby DE1 2QY, UK.
7 Departments of Clinical Medicine and Immunology, Institute of Molecular Medicine, Trinity College Dublin, Dublin, Ireland.
8 Wellcome Trust Centre for Human Genetics and
9 Gastroenterology Unit, University of Oxford, Oxford OX3 7BN, UK.
10 Gastroenterology Section, Imperial College London, Hammersmith Hospital, London W12 0HS, UK.
11 Department ofGastroenterology and Liver Unit, Royal Hallamshire Hospital, Sheffield S10 2JF, UK.
12 Department of Gastroenterology, St. James's University Hospital, Leeds LS9 7TF, UK.
13 Department of Gastroenterology, Llandough Hospital, Penarth CF64 2XX, UK.
14 Department of Paediatric Gastroenterology, Leiden University Medical Center, 2300 RC Leiden, The Netherlands.
15 Department of Pediatric Gastroenterology and
16 Department of Gastroenterology, Vrije University Medical Center, 1007 MB Amsterdam, The Netherlands.
§ These authors contributed equally to this work.

Summary

We tested 310,605 SNPs for association in 778 individuals with celiac disease and 1,422 controls. Outside the HLA region, the most significant finding (rs13119723; $P = 2.0 \times 10^{-7}$) was in the *KIAA1109-TENR-IL2-IL21* linkage disequilibrium block. We independently confirmed association in two further collections (strongest association at rs6822844, 24 kb 5' of IL21; meta-analysis $P = 1.2 \times 10^{-14}$, odds ratio = 0.63), suggesting that genetic variation in this region predisposes to celiac disease.

Introduction

Celiac disease is a common (1% prevalence) small intestinal inflammatory condition induced by dietary wheat, rye and barley. However, despite high heritability (estimated at 87% from twin studies[1]), no non-HLA genetic risk factors have been identified and convincingly replicated. The majority of individuals with celiac disease possess HLA-DQ2 (and the remainder mostly HLA-DQ8)[2], and the mechanism by which HLA-DQ2 presents cereal peptides to intestinal T cells is understood[3]. However, HLA-DQ2 is common in healthy individuals, demonstrating that it contributes to, but is not sufficient for, disease development.

Therefore, we designed a genome-wide association (GWA) study to identify predisposing genetic factors in celiac disease. We genotyped samples with Illumina BeadChips (Supplementary Methods). After quality control, we performed association analysis on 310,605 SNPs with minor allele frequency >1% genotyped in 778 UK individuals with celiac disease and 1,422 UK population controls (Supplementary Table 1). The overall SNP call rate was 99.87% (see Supplementary Fig. 1 for single-SNP association statistics). We saw highly significant association around the HLA locus, as expected. Association was strongest at rs2187668, which maps to the first intron of HLA-DQA1 ($\chi^2 = 769.1$, $P < 10^{-19}$; frequency of A allele in controls, 13.8%; affected individuals, 53.1%; odds ratio (OR) = 7.04 (95% confidence interval (c.i.) 6.08–8.15)). When compared with classical HLA typing (Supplementary Methods), the rs2187668-A allele tagged HLA-DQ2.5*cis* efficiently ($r^2 = 0.97$, Supplementary Table 2). HLA-DQ2.5*cis,* in which the two chains of the DQ2 heterodimer are encoded on the same chromosome, is the most common HLA-DQ2 haplotype associated with celiac disease. One or two copies of HLA-DQ2.5*cis* (inferred by rs2187668 genotype) were present in 89.2% of UK participants with celiac disease versus 25.5% of population controls. To identify other HLA predisposing variants occurring in the presence, or absence, of HLA-DQ2.5*cis,* we performed further analyses stratified by rs2187668 genotype. In affected individuals (n = 558) and in controls (n = 331) of rs2187668-AG genotype, we saw peak association at rs9357152 ($P = 5.2 \times 10-14$); in affected individuals (n = 83) and controls (n = 1,059) of rs2187668-GG genotype, we saw peak association at rs9275141 ($P = 3.9 \times 10^{-16}$). There were too few rs2187668-AA cases (n = 31) for analysis. The finding that rs2187668, rs9275141 and rs9357152 map within or adjacent to HLA-DQA1 and HLA-DQB1 underscores the critical role of HLA-DQ2/8 in antigen presentation in celiac disease.

Outside the HLA region, we observed a greater number of significantly associated SNPs than would be expected by chance, with 56 SNPs showing association at $P = 10^{-4}$ (Supplementary Table 3). Many of these SNPs are in close proximity, suggesting that some of the excess in SNPs with low P values might be due to true disease associations among multiple SNPs in linkage disequilibrium (LD) with nearby disease variants. Therefore, we prioritized these findings for rapid replication (Supplementary Table 3 shows interim results) while designing a more extensive SNP replication study. We noted weak evidence for association in the previously reported *CD28-CTLA4-ICOS* region4 (rs4675374, P = 0.007; rs11681040, P = 0.008) but not the MYO9B region[5].

The most significant (non-HLA) finding was rs13119723 ($P = 2.0 \times 10^{-7}$; frequency of G allele in controls = 15.8%; affected individuals, 10.1%). Permutation of affection status labels demonstrated genomewide significance: in 9 of 200 (P = 0.045) permutations, the most significant permuted P value was 2.0×10^{-7}. The location of rs13119723 close to IL2 and IL21 made it a highly plausible celiac disease candidate gene. We did not observe any evidence for statistical interaction between rs13119723 genotype and inferred HLA-DQ2.5cis genotype (P = 0.20). We then confirmed association of rs13119723 with celiac disease in two separate collections (Table 1). The G allele of rs13119723 was more common in controls in each collection, and meta-analysis (of all 4,680 samples) established highly significant disease association at rs13119723 ($P = 4.8 \times 10^{-11}$).

Table 1

Chromosome 4q27 markers in the UK genome-wide association scan and replication studies

	UK GWA scan collection Allele frequency (%)			Dutch collection Allele frequency (%)		
	Cases n = 778	Controls n = 1,422	P [a]	Cases n = 508	Controls n = 929	P [a]
rs6835946	30.3	29.6	0.63	25.1	25.5	0.81
rs11938795	21.2	26.3	0.00017			
rs4374642	10.5	8.4	0.020	12.1	8.4	0.0015
rs13151961	12.6	17.9	5.2×10^{-6}	12.2	19.1	2.2×10^{-6}
rs4505848	34.4	32.3	0.15	37.6	34.9	0.15
rs4288027	10.4	8.4	0.02			
rs7683061	40.6	38.0	0.090			
rs13119723	10.1	15.8	2.0×10^{-7}	11.5	16.4	0.00042
rs11734090	21.5	26.5	0.00023			
rs7699742	34.9	32.2	0.066			
rs1127348	24.8	21.6	0.016	25.7	24.0	0.34
rs7678445	9.9	7.7	0.013			
rs7684187	25.2	30.1	0.00056			
rs11732095	8.8	8.6	0.79	9.9	9.1	0.45
rs716501	34.8	32.2	0.087			
rs10857092	5.2	6.6	0.069	7.8	7.3	0.60
rs6848139	10.2	8.5	0.050			
rs6852535	29.7	29.4	0.84			
rs12642902	28.5	34.6	3.3×10^{-5}			
rs6822844	12.6	17.9	4.6×10^{-6}	12.4	18.5	2.1×10^{-5}
rs4492018	26.6	26.5	0.93	22.7	23.1	0.80
rs975405	38.6	42.8	0.0071	41.5	43.3	0.34
rs7682241	36.6	34.4	0.13			
rs17005931	26.5	26.2	0.82			
rs1398553	34.1	30.6	0.016	35.2	32.6	0.18
rs2893008	10.0	8.0	0.021	11.0	7.4	0.0011
rs6840978	16.3	21.5	3.8×10^{-5}	15.2	21.9	2.0×10^{-5}

Written consent was obtained from all participants. The study was approved by Oxfordshire REC B, the Medical Ethical Committee of the University Medical Center Utrecht and the Institutional Ethics Committee of St. James's Hospital. Boldface indicates most significant SNPs overall.

[a] P values from χ^2 test of allele counts. All tests are two tailed.

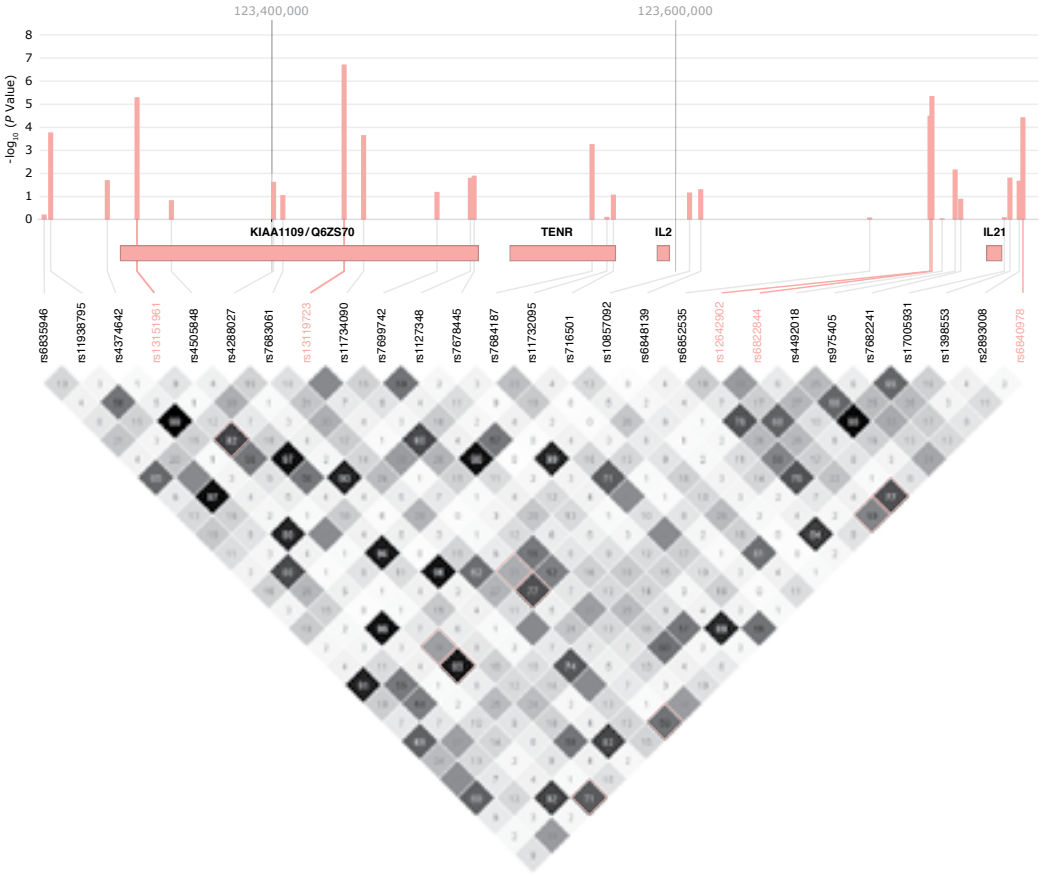
Irish collection				
Allele frequency (%)			Meta-analysis	
Cases n = 483	Controls n = 560	P [a]	OR (95% CI)	P [a]
31.4	33.5	0.32		
8.3	6.7	0.17		
14.8	19.4	0.0056	0.65 (0.58–0.73)	1.3×10^{-12}
31.3	27.8	0.084		
12.8	16.2	0.030	0.66 (0.58–0.74)	4.8×10^{-11}
21.7	16.9	0.0053		
8.5	7.7	0.50		
7.0	7.7	0.53		
14.2	19.7	0.0013	0.63 (0.57–0.71)	1.3×10^{-14}
29.1	30.2	0.60		
39.9	43.9	0.070		
30.6	25.7	0.015		
7.3	6.5	0.50		
19.2	24.0	0.0083	0.70 (0.63–0.78)	1.1×10^{-10}

rs13119723 maps to a region of strong LD (Supplementary Fig. 2). In our original scan, we genotyped 27 SNPs in this 4q27 region, covering B480 kb from rs6835946 to rs6840978. In addition to rs13119723, four other SNPs showed association with celiac disease at $P = 10^{-4}$ in the UK data set (Fig. 1 and Table 1). We further genotyped rs6822844, rs13151961 and rs6840978 (all strongly correlated with rs13119723; Fig. 1) in the Dutch and Irish collections and replicated the associations observed in the UK data set (Table 1 and Supplementary Table 4).We observed the strongest association overall at rs6822844, approximately 24 kb 5' of IL21 (meta-analysis $P = 1.3 \times 10^{-14}$, OR = 0.63 (0.57–0.71)). Markers on the HumanHap300 BeadChip (Illumina) are haplotype tag SNPs. We found that the 27 SNPs genotyped in the UK collection very efficiently captured the common genetic variation in the ~480-kb region (161 of 165 common phase I+II HapMap SNPs pairwise tagged at $r^2 = 0.8$ in CEU population[6]; Supplementary Methods). Therefore, genotyping of further markers in the UK collection was unlikely to contribute substantial additional information.

Finer analysis of haplotype structure in the ~480-kb region in the UK collection showed subdivision into two closely correlated ~439-kb and ~40-kb haplotype blocks (using strict criteria[7]). We found the rs13119723-G allele on a single strongly associated haplotype in both blocks (Supplementary Fig. 2), with haplotype frequencies of 10.1% in affected individuals and 15.3% in controls in the 439-kb block ($P = 2.1 \times 10^{-6}$) and 16.3% in affected individuals and 21.5% in controls in the 40-kb block ($P = 4.3 \times 10^{-5}$). We genotyped ten additional SNPs to tag haplotypes of frequency 45% (in addition to the four SNPs already tested) in the Dutch and Irish collections and found similar haplotype structure and association across all three populations (Supplementary Table 4). Because of extensive LD, these analyses did not allow us to determine the causal variant associated with celiac disease in the 4q27 region. The population-specific genetic variance at the associated 4q27 markers (CEU HapMap data) is relatively high, suggesting possible selection in the Northern European population.

Figure 1

Analysis of chromosome 4q27 region around rs13119723. A ~480-kb region between rs6835946 and rs6840978 is shown (build 36 map), with single-SNP allelic association test P values, genes and LD statistics (r^2) determined from the UK data set.



The 4q27 celiac disease–associated region contains three known protein-coding genes (*TENR* (also known as *NM_139243*; official gene symbol pending), *IL2* and *IL21*) and a predicted gene of unknown function (*KIAA1109*). We manually annotated the human genome sequence in the region (data not shown) but did not identify further genes. IL-2, secreted in an autocrine fashion by antigen-stimulated T cells, is a key cytokine for T cell activation and proliferation. Another T cell–derived cytokine, IL-21, enhances B, T and NK cell proliferation and interferon-γ production. Both cytokines are implicated in the mechanisms of other intestinal inflammatory diseases[8,9]. We examined expression profiles for the four genes across multiple cell and tissue types in the GNF SymAtlas database (Supplementary Methods). *TENR* is specifically expressed in testis and is an unlikely candidate for the causal celiac disease susceptibility gene. The function of KIAA1109 is largely unknown[10], although *KIAA1109* is widely expressed as multiple splice variants in multiple tissues. We looked specifically at gene expression in duodenal tissue from normal individuals and those with celiac disease (with normal histology or with villous atrophy). *TENR* expression was mostly undetectable. We did not see any differences between normal individuals and those treated for celiac disease for *KIAA1109*, *IL2* or *IL21*. In the presence of inflammation (in individuals with untreated celiac disease), *KIAA1109* and *IL2* levels showed a modest reduction, and *IL21* showed an increase (Supplementary Fig. 3). The region in mouse syntenic to human 4q27 (*Idd3*) determines susceptibility to multiple autoimmune diseases in the NOD mouse model by a mechanism influencing IL2 mRNA and IL-2 protein levels and CD4+ CD25+ regulatory T cell activity[11]. However, further studies are required to determine the human gene affecting susceptibility to celiac disease in this region.

Our GWA study has identified genetic variation in an LD block encompassing the *KIAA1109-TENR-IL2-IL21* genes as a new susceptibility factor for celiac disease. In addition to further investigation of this 4q27 region, the next steps in dissecting the genetic causes of celiac disease include larger-scale replication of other putative associations and additional genome-wide analyses (for example, of copy number variation[12]).

Acknowledgments

We thank C.A. Mein and thwve Barts and The London Genome Centre for advice and genotyping support; D. Simpkin, T. Dibling and C. Hand for genotyping (Sanger Institute); M.J. Caulfield for advice on study design; D.P. Kelsell for comments on the manuscript; D. Strachan and W.L. McArdle for 1958 Birth Cohort samples; A. Monsuur for patient recruitment; G. Meijer and J. Meijer for histology review; K. Duran for DNA extraction; H. van Someren for clinical database management (The Netherlands); A. Ryan, G. Turner, M. Abuzakouk, N. Kennedy, F. Stevens and C. O'Morain for patient and control recruitment and sample management (Ireland). We thank J. Loveland for EST annotation and checking. We thank the Wellcome Trust Centre for Human Genetics, University of Oxford, for computing facilities. We thank all affected individuals and controls for participating in this study. We acknowledge funding from Coeliac UK; the Coeliac Disease Consortium (an innovative cluster approved by The Netherlands Genomics Initiative and partly funded by the Dutch government (grant BSIK03009)); The Netherlands Genomics Initiative (grant 050-72-425); The Netherlands Organization for Scientific Research (grant 901-04-219); the Science Foundation Ireland and the Wellcome Trust (GR068094MA Clinician Scientist Fellowship to D.A.v.H.; New Blood Fellowship to R.M. and support for the work of R. McG. and P.D.). The authors acknowledge use of DNA from the British 1958 Birth Cohort collection, funded by the UK Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02.

References

1 Nistico, L. *et al.* Gut 55, 803–808 (2006).
2 Karell, K. *et al.* Hum. Immunol. 64, 469–477 (2003).
3 van Heel, D.A. & West, J. Gut 55, 1037–1046 (2006).
4 Hunt, K.A. *et al.* Eur. J. Hum. Genet. 13, 440–444 (2005).
5 Monsuur, A.J. *et al.* Nat. Genet. 37, 1341–1344 (2005).
6 International HapMap Consortium. Nature 437, 1299–1320 (2005).
7 Gabriel, S.B. *et al.* Science 296, 2225–2229 (2002).
8 Sadlack, B. *et al.* Cell 75, 253–261 (1993).
9 Monteleone, G. *et al.* Gastroenterology 128, 687–694 (2005).
10 He, Q.Y. *et al.* Bioinformatics 22, 2189–2191 (2006).
11 Yamanouchi, J. *et al.* Nat. Genet. 39, 329–337 (2007).
12 Redon, R. *et al.* Nature 444, 444–454 (2006).

Supplementary Material

UK subjects

Celiac disease patients were recruited from adult outpatient clinics at seven UK hospital sites (Barts and the London, London; Derbyshire Royal Infirmary, Derby; Hammersmith Hospital, London; John Radcliffe Hospital, Oxford; Leeds University Hospitals, Leeds; Llandough Hospital, Cardiff; Sheffield University Hospitals, Sheffield). Inclusion criteria were as described[1], based on presence of villous atrophy at diagnosis and (since test introduction) positive anti¬endomysial/tissue transglutaminase antibody (Supplementary Table 1). Population-based controls were analysed from the 1958 British Birth Cohort. Ethics committee (Oxfordshire REC B) and local approval were obtained for all cohorts. Genomic DNA was extracted from peripheral blood, or from immortalised peripheral blood lymphocyte cell lines (1958 British Birth Cohort cohort). All individuals were unrelated and of white northern European ethnic origin.

Dutch subjects

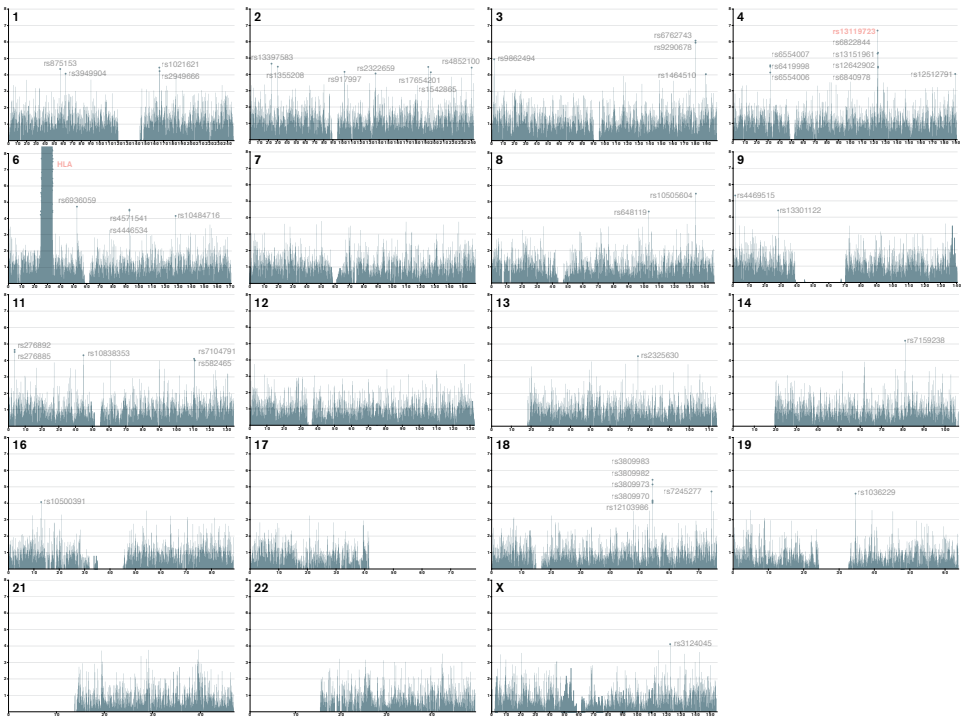
DNA, isolated from whole blood, was obtained from unrelated Dutch individuals with celiac disease. All the affected individuals were diagnosed in accordance with the revised ESPGAN criteria[2]. More than 90% of the affected individuals were HLA-DQ2-positive. The initial biopsy specimens of the individuals were retrieved; all showed a Marsh III lesion upon re-evaluation by one of two experienced pathologists. Both cohorts included children and adults. The control cohort comprised unrelated individuals that were random blood bank donors. All cases and controls were from The Netherlands and of European descent, and at least three of their four grandparents were also born in The Netherlands. This study was approved by the Medical Ethical Committee of the University Medical Center Utrecht.

Irish subjects

Celiac disease patients were recruited at St. James's Hospital and the Adelaide and Meath (AMiNCH) Hospitals in Dublin, and University College Hospital, Galway, Ireland. Patients were diagnosed on the basis of histological appearance, antibody positivity and clinical improvement in response to a gluten free diet. Over 90% of patients have histological lesions with a Marsh II or Marsh III classification. Patients with Marsh I (intra epithelial lymphocyte infiltrates) were only diagnosed as celiac disease if they have both positive EMA/anti-tTG anti-

Supplementary Figure 1

Summary of genome wide association scan association results. Y axis scale on each chromosome graph is $-\log_{10}$ P value (allele count χ^2 test, two tailed). SNPs with observed $P<10^{-4}$ are annotated with reference SNP number.



10,000,000 bp

Deletion

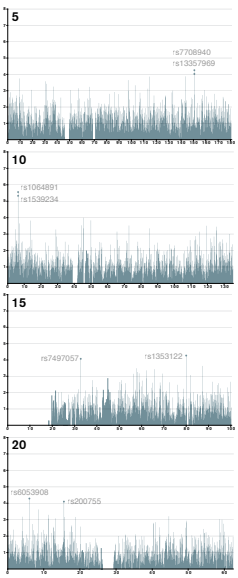
body tests and clinical improvement in response to a gluten free diet. Unselected population controls were collected anonymously from blood transfusion donors and healthy volunteers. Ethics committee (Institutional Ethics Committee of St James's Hospital) and local approval were obtained for all cohorts. All individuals were unrelated and of white northern European ethnic origin.

HLA typing

A subset of UK samples (n=266 of final dataset) was HLA genotyped at the Transplant Immunology laboratory, Oxford Radcliffe Hospitals NHS Trust. A set of 56 sequence specific reactions across HLA-DQA1 and HLA-DQB1 was tested using a PCR based method[3]. External quality control was provided by United Kingdom National External Quality Assessment for Histocompatibility and Immunogenetics
and by exchange of reference samples under the University of California at Los Angeles (UCLA) Immunogenetics Center International Cell Exchange program.

Genome wide association scan genotyping

Genotyping was performed according to the Infinium II protocol from Illumina (Illumina, San Diego, USA). DNA concentration was measured by picogreen assay. In brief, 750ng of genomic DNA was whole genome amplified, fragmented, hybridised to HumanHap300 BeadChip (cases) or HumanHap550 BeadChip (controls), extended, stained and imaged. The HumanHap550 BeadChip contains identical assays and probes (in the same beadpools) as the HumanHap300 BeadChip, but with additional SNP assays. Samples were only included if a minimum 95% call rate was observed for the sample in Beadstudio across the 313,505 SNPs common to both HumanHap300 and HumanHap550 BeadChips. For each SNP assay, normalised R and theta values were exported from BeadStudio. SNP R and theta data from either HumanHap300 or HumanHap550 assay were then combined, and genotypes clustered by a novel algorithm.



Genome wide association scan calling algorithm

The Illumina Infinium II assay uses two allele specific Cy3/Cy5 fluorescent probes per SNP. When investigating all samples three clusters are visible when plotting the intensity of allele B against the intensity of allele A. Data from large sample sets (here, n=2200) is helpful in more accurately assigning these clusters. BeadStudio v2.3.42 was unable to simultaneously call genotypes from all samples, because different BeadChips had been used for the controls and cases, even though actual beadpools were identical.

To overcome potential spurious associations due to differences in this calling of cases and controls, an algorithm was developed which calls genotypes for each SNP, while using all raw data. It relies on a measure (*theta*) per individual (*i*), which is defined as:

$$\text{theta}_i = \frac{2}{\pi}\arctan\left(\frac{\text{AlleleB}_i}{\text{AlleleA}_i}\right)$$

For each SNP, three clusters were defined, representing the homozygous AA genotypes (individuals with a low *theta* value), heterozygous AB genotypes (individuals with an intermediary *theta* value) and homozygous BB genotypes (individuals with high *theta* values).

In order to identify the appropriate clusters first a scoring function was defined which allowed for determining how well the three clusters had been defined. For the heterozygous cluster the variance was calculated:

$$\sigma_{AB}^2 = \sum_{i=1}^{i=n_{ab}} (\text{theta}_i - \mu_{AB})^2$$

Where *i* = each individual in cluster AB, *nab* = number of individuals in this cluster, μABi = mean *theta* value for all individuals for SNP *i* in this cluster

For the homozygous AA cluster a deviation from the expected minimal *theta* value ($theta_{min}$) was defined:

$$\sigma_{AA}^2 = \sum_{i=1}^{i=n_{aa}} (\text{theta}_i - \text{theta}_{\text{min}})^2$$

Where *i* = each individual AA, and n_{aa} = number of individuals in this cluster.

For the homozygous BB cluster a deviation from the expected maximal *theta* value ($theta_{max}$) was defined:

$$\sigma_{BB}^2 = \sum_{i=1}^{i=n_{bb}} (\text{theta}_i - \text{theta}_{\text{max}})^2$$

Where *i* = each individual BB, and n_{bb} = number of individuals in this cluster.

The total variance (σ^2) was determined by summing the three independent variances:

$$\sigma^2 = \sigma_{AA}^2 + \sigma_{AB}^2 + \sigma_{BB}^2$$

Subsequently the entire search space was assessed, and the three clusters with the lowest total variance were considered to represent the correct genotypes.

Once for each individual initial genotypes were assigned, it was determined per individual whether its theta value corresponded well to the cluster it had designated, or that it was also likely it would fit in another cluster. When an individual had a *theta* value that had a squared distance to the mean of its own cluster that was more than the half the squared distance to the mean of another cluster, the initially assigned genotype for this individual was removed.

Java source code, implementing this algorithm, is available at www.prioritizer.nl/wga.php

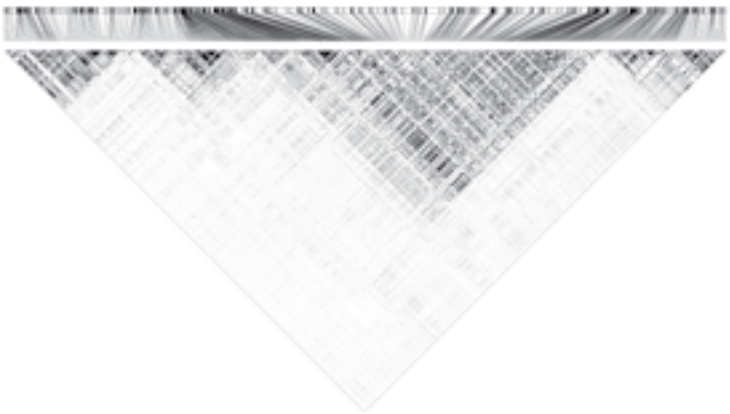
Supplementary Figure 2

Linkage disequilibrium and haplotype analysis of 4q27 region

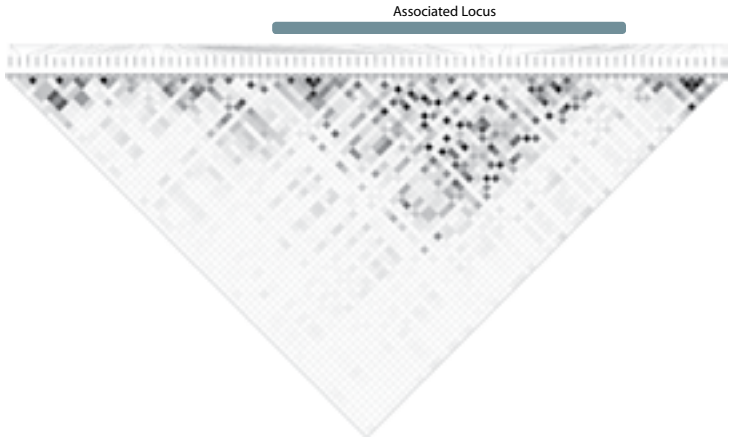
a) Haploview linkage disequilibrium analysis for 1Mb region around rs13119723. Region from rs6825926 to rs309375 analysed. HapMap phase I+II data shown for 539 SNPs with >95% call rate and >1% mean allele frequency in 30 CEU trios. UK coeliac and population control data shown for 81 SNPs in 2200 samples. Greyscale boxes indicates pairwise correlation statistic (r^2). Red bar indicates ~480kb celiac disease associated region. **b)** Haplotype analysis of ~480kb celiac disease associated region in UK dataset. Region from rs6835946 to rs6840978 analysed. Haplotypes of >1% frequency (determined in Haploview using Gabriel *et al* criteria) shown. Connections across the two blocks shown for haplotypes >10% frequency with thick lines, and >1% frequency with thin lines. Multiallelic D' statistics between the blocks are shown.

a)

HapMap CEU Dataset
(n = 30 trios)



Celiac UK Dataset
(n = 2,200 unrelated individuals)



b)

Block 1

rs6835946	rs11938795	rs4374642	rs13151961	rs4505848	rs4288027	rs7683061	rs13119723	rs11734090	rs7699742	rs1127348	rs7678445	rs7684187	rs11732095	rs716501	rs10857092	rs8848139	rs6852535	rs12642902	rs6822844	Case Freq	Control Freq	P value
1	1	1	1	1	1	1	1	1	1	1	2	1	1	2	2	1	1	2	2	0.294	0.287	0.62
2	1	1	1	2	1	2	1	1	2	2	2	1	1	1	2	1	2	2	2	0.232	0.204	0.031
2	**2**	**1**	**2**	**1**	**1**	**2**	**2**	**2**	**1**	**1**	**2**	**2**	**1**	**2**	**2**	**1**	**2**	**1**	**1**	**0.101**	**0.153**	**2.1x10^{-6}**
2	1	2	1	1	2	1	1	1	1	1	1	1	1	2	2	2	2	2	2	0.099	0.077	0.015
2	2	1	1	1	1	2	1	2	1	1	2	2	2	2	2	1	2	1	2	0.080	0.072	0.31
2	1	1	1	2	1	2	1	1	2	1	2	1	1	1	2	1	2	2	2	0.057	0.050	0.28
2	1	1	1	2	1	2	1	1	2	1	2	1	1	1	1	1	2	1	2	0.040	0.047	0.30
2	1	1	1	1	1	2	1	1	1	1	2	2	1	2	2	1	2	1	2	0.034	0.035	0.93
2	2	1	2	1	1	2	1	2	1	1	2	2	1	2	2	1	2	1	1	0.022	0.022	0.97

Block 2

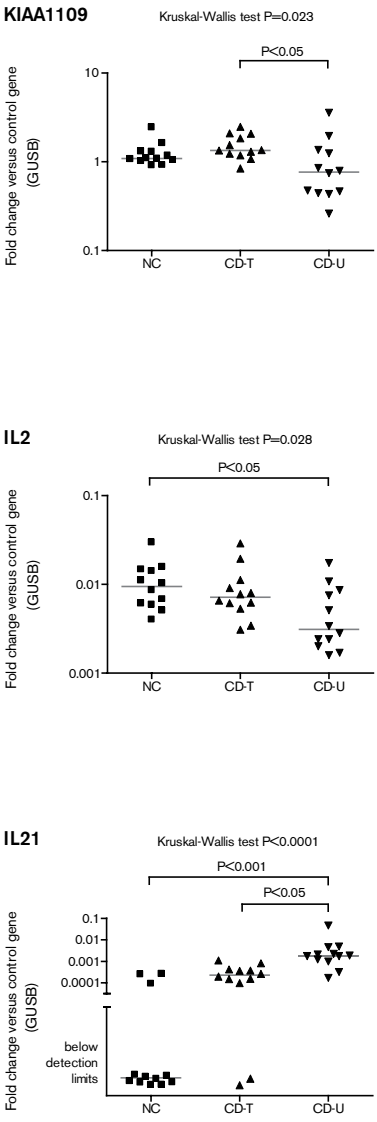
rs4492018	rs975405	rs7682241	rs17005931	rs1398553	rs2893008	rs6840978	Case Freq	Control Freq	P value
2	1	2	2	1	1	2	0.341	0.306	0.017
1	1	1	1	2	1	2	0.266	0.263	0.84
2	**2**	**2**	**2**	**2**	**1**	**1**	**0.163**	**0.215**	**4.3x10^{-5}**
2	2	2	2	2	1	2	0.123	0.134	0.32
2	2	1	2	2	2	2	0.101	0.080	0.019

D'0.95

Supplementary Figure 3

Quantitative RT-PCR analysis of 4q27 region gene expression in duodenal tissue

Fold change ($2^{-\Delta\Delta Ct}$) versus control gene (GUSB) shown. Median shown with a bar. NC, normal control individuals (n=12); CD-T treated coeliac individuals with normal duodenal histology (n=12); CD-U untreated coeliac individuals with villous atrophy (active inflammation) observed on duodenal histology (n=12).



Genome wide association scan quality control and statistical analysis

Samples were then excluded for the following reasons: gender inferred by X chromosome genotype different to clinical record of gender (n=8), possible ethnic outliers (n=3) identified by PLINK ethnic outlier (nearest neighbour) analysis, deliberate and inadvertent duplicates identified by whole genome identity by state analysis (PLINK), mismatch with >1 SNP previously genotyped by another method (n=9).

Of the 313,505 SNPs common to the Human Hap300 and HumanHap550 BeadChips, we mapped 312,867 SNPs to autosomes or the X chromosome in NCBI build 36, and selected these for further analysis. SNPs were then excluded from analysis for the following reasons (applied in order): call rate <95% in either case or control samples (n=1,490), allele frequency <1% in combined samples (n=32), deviation from Hardy Weinberg equilibrium (P<0.0001 in control samples, n=740). Subsequent analyses were performed on a dataset of 310,605 SNPs with a call rate of 99.87%. No difference in call rates was seen between cases (Hap300 BeadChip) and controls (Hap550 BeadChip).

A previous study had reported association of celiac disease with genetic variation in *MYO9B*[4]. Six SNPs reported in the previous study were tested in the current genome scan dataset (rs7246865, rs4808571, rs1870068, rs10409451, rs2305767, rs962917).

The genome scan was 80% powered to detect an allelic association which had a $P < 0.001$, with the sample size we used, and assuming a MAF of 0.26 in controls (the mean SNP allele frequency in controls of all SNPs tested) with an odds ratio of 1.22.

Association analysis at single marker level was performed by chi-squared test of allele counts as implemented in PLINK v0.99q software provided by S. Purcell and colleagues (pngu.mgh.harvard.edu/~purcell/plink/). Logistic regression implemented in R statistical software (www.r-project.org) was applied to calculate odds ratios and confidence intervals and to test for possible differences in disease susceptibility due to statistical interaction between HLA and non-HLA genotypes. In some analyses the broad HLA region (defined as 20,000,000 to 39,999,999 base pairs on chromosome 6 NCBI Build 36 map) was excluded (3,226 SNPs in final dataset).

Haploview v3.32 was used for linkage disequilibrium block characterisation, and haplotype association statistics[5]. Tagging efficiency of the 27 Illumina SNPs in the 4q27 region, versus HapMap data (Release 21a, >5% frequency HapMap phase I+II CEU SNPs with >95% genotype data), was assessed using Tagger in Haploview[6], with a pairwise tagging approach.

Supplementary table 1

Characteristics of subjects

	Genome wide association scan		Replication	
	UK celiac cases	UK population controls	Dutch celiac cases	Dutch healthy controls
Individuals (n)	778	1422	508	929
Female (%)	71.9%	49.4%	68.8%	38.7%
Median age at diagnosis, range (years)	43 (0.3 - 84)	Born in 1958 (49 - 49)	46.3 (1-83)	Not available
Intestinal histology when untreated	100% villous atrophy (Marsh III)	-	100% villous atrophy (Marsh III)	-
Known positive IgA anti-endomysial / tTG anti-body when untreated	71.4%	-	Not available	-

Supplementary table 2

Classical HLA type and rs2187668 genotype.

UK celiac cases (n=246)	**rs2187668 genotype**		
	AA	AG	GG
Classical HLA type			
DQ2.5*cis*[1] homozygous	49	2	0
DQ2.5*cis* heterozygous	1	167	1
DQ2.5*trans*[2], DQ8 homozygous or DQ8 heterozygous but not DQ2.5*cis*	0	0	26

[1]DQ2.5*cis* (DQA1*0501 and DQB1*0201 encoded on the same chromosome)
[2]DQ2.5*trans* (DQA1*0505 and DQB1*0202 encoded on different chromosomes)

This table does not include n=17 celiac cases heterozygous for HLA-DQ2.5*cis* in addition to other rare DQA1/DQB1 alleles (all were heterozygous or homozygous for rs2187668 A), nor 3 individuals encoding a partial DQ2 heterodimer.

Replication	
Irish celiac cases	Irish healthy controls
483	560
71.0%	51.0%
Not available	Not available
>90% crypt hyperplastic lesion, villous atrophy (Marsh II/III)	-
>90%	-

As a final quality control check, we visually inspected genotype clouds for SNPs with single marker association test $P<10^{-3}$. Apparent association ($P=5.8 \times 10^{-11}$) with SNP rs149947 was spurious due to poor genotype cloud clustering, this SNP was removed from analyses.

Genome wide association scan genotyping data are available from https://enigma.sanger.ac.uk/cgi-bin/PostGenomics/genotyping/manager. Please follow instructions to obtain a login/password which will be made available to bona-fide scientific investigators.

Replication collections: genotyping and statistical analysis

We genotyped 8 SNPs outside the 4q27 region, and 14 SNPs from the 4q27 region, in further independent cohorts (Table 1, Supplementary Table 3, 4). An additional 4q27 region SNP rs4833837 (synonymous SNP in *IL21*) was tested in the Dutch collection only (Supplementary Table 4). A collection of Dutch healthy controls (blood donors) and Dutch celiac cases was genotyped at University Medical Centre Utrecht. Irish celiac case and control collections were genotyped at Queen Mary University of London. Taqman assays (Applied Biosystems, Warrington, UK) were used for genotyping at both sites. Association analysis was performed by chi-squared test of allele counts, and meta-analysis of all cohorts by the Maentel-Haenszel method.

We observed perfect concordance between rs13119723 genotypes in n=123 Dutch controls genotyped by both Taqman and Hap300 BeadChip Infinium assay (as part of a separate project).

Annotation of transcripts and features in the 4q27 region.

Manual annotation of coding and non-coding transcript structures was completed using the Sanger HAVANA analysis and annotation pipeline (www.sanger.ac.uk/HGP/havana/) on a 559 kb region of chromosome 4:123,280,304-123,840,222 (NCBI36, March 2006).

The following features were observed: 1. Known_CDS AC022489.2 (aka KIA1109, FSA); 2. Known_CDS, AC022489.3 novel protein, possible orthologue of mouse testis nuclear RNA-binding protein (Tenr)[3]. Known_CDS IL2; 4. Known_CDS IL21 (non-coding novel_transcript annotated AC053545.3 on opposite strand overlapping with IL21 exon 1).

No known copy number variants (detected by manual inspection of SNP fluorescence intensity plots, or recorded in the Database of Genomic Variants: projects.tcag.ca/variation/) or microRNA's were observed in the region.

Gene expression analysis

Expression of genes in the 4q27 region across multiple human tissues was examined using the HumanGeneAtlas GNF1H qcRMA dataset in the GNF SymAtlas v1.2.4 database (symatlas.gnf.org/SymAtlas/)[7]. Similar patterns for *KIAA1109*, suggesting broad expression across multiple tissues, were observed in the GeneHub-GEPIS database based on EST analysis (www.cgl.ucsf.edu/Research/genentech/genehub-gepis/index.html).

Collection of duodenal biopsies, RNA extraction, cDNA synthesis, and Taqman quantitative RT-PCR using Applied Biosystems pre-designed assays was performed as described[8]. Each sample/condition was assayed in triplicate. Fold change ($2^{-\Delta\Delta Ct}$) versus control gene (*GUSB*) was calculated. *IL21* expression levels were undetectable in at least two of the triplicate assays for some samples, therefore we set Ct for the sample to an arbitrary high value. Median and Kruskal-Wallis test statistics were calculated, with Dunn's Multiple Comparison tests (Supplementary Fig. 3).

References supplementary material

1 Hunt, K.A. *et al.* Eur J Hum Genet 13, 440-4 (2005).
2 Eur J Gastroenterol Hepatol 13, 1123-8 (2001).
3 Bunce, M. *et al.* Tissue Antigens 46, 355-67 (1995).
4 Monsuur, A.J. *et al.* Nat Genet 12, 1341-4 (2005).
5 Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Bioinformatics 21, 263-5 (2005).
6 de Bakker, P.I. *et al.* Nat Genet 37, 1217-23 (2005).
7 Su, A.I. *et al.* Proc Natl Acad Sci U S A 101, 6062-7 (2004).
8 Diosdado, B. *et al.* Clin Gastroenterol Hepatol (2007).

Supplementary table 3

Genome wide association and replication results

a) Genome wide association results for SNPs with $P<10^{-4}$ (excluding HLA region)

SNP	Chr	Position	UK genome wide scan					Nearest gene	Replicated in Dutch + Irish
			HWE Controls	MAF Controls	MAF Cases	Odds Ratio [95% CI]	P value		
rs875153	1	56672858	0.368	0.494	0.43	0.77 [0.68 - 0.87]	4.3×10^{-5}	PPAP2B	
rs3949904	1	62620678	0.518	0.287	0.345	1.30 [1.14 - 1.49]	8.5×10^{-5}	USP1	
rs1021621	1	165465160	0.223	0.491	0.425	0.77 [0.68 - 0.87]	3.6×10^{-5}	POU2F1	No (Table S3B)
rs2949666	1	165659326	0.222	0.48	0.417	0.77 [0.68 - 0.88]	5.7×10^{-5}	POU2F1	
rs13397583	2	23459535	0.732	0.367	0.303	0.75 [0.66 - 0.86]	2.2×10^{-5}	UBXD4	
rs1355208	2	30298826	0.642	0.351	0.415	1.31 [1.15 - 1.48]	3.4×10^{-5}	LBH	
rs917997	2	102437000	0.027	0.215	0.268	1.34 [1.16 - 1.54]	6.8×10^{-5}	IL18RAP	
rs2322659	2	136272129	0.602	0.188	0.238	1.35 [1.16 - 1.57]	8.6×10^{-5}	LCT	
rs17654201	2	193432325	0.266	0.063	0.034	0.52 [0.38 - 0.71]	3.5×10^{-5}	TMEFF2	
rs1542865	2	196017781	0.574	0.171	0.126	0.70 [0.58 - 0.83]	7.2×10^{-5}	SLC39A10	
rs4852100	2	240236764	0.343	0.239	0.296	1.34 [1.17 - 1.54]	3.8×10^{-5}	NP_997366.1	
rs9862494	3	2377673	0.723	0.04	0.07	1.82 [1.39 - 2.39]	1.1×10^{-5}	CNTN4	
rs6762743	3	180494694	0.749	0.293	0.224	0.70 [0.60 - 0.80]	8.1×10^{-7}	ZNF639	No (Table S3B)
rs9290678	3	180499714	0.749	0.292	0.224	0.70 [0.61 - 0.81]	1.1×10^{-6}	ZNF639	No (Table S3B)
rs1464510	3	189595248	0.261	0.457	0.519	1.28 [1.13 - 1.45]	9.0×10^{-5}	LPP	
rs6554006	4	31769556	0.777	0.104	0.068	0.63 [0.50 - 0.79]	7.6×10^{-5}	FCDH7	
rs6554007	4	31769569	0.411	0.263	0.206	0.73 [0.63 - 0.84]	2.7×10^{-5}	FCDH7	
rs6419998	4	31783944	0.369	0.292	0.233	0.74 [0.64 - 0.85]	3.3×10^{-5}	FCDH7	
rs13151961	4	123334952	0.588	0.179	0.126	0.66 [0.55 - 0.79]	5.2×10^{-6}	KIAA1109	
rs13119723	4	123437763	0.193	0.158	0.101	0.60 [0.50 - 0.73]	2.0×10^{-7}	KIAA1109	
rs12642902	4	123727951	0.598	0.346	0.285	0.75 [0.66 - 0.86]	3.3×10^{-5}	IL21	Yes (Table 1)
rs6822844	4	123728871	0.588	0.179	0.126	0.66 [0.55 - 0.79]	4.6×10^{-6}	IL21	Yes (Table 1)
rs6840978	4	123774157	0.529	0.215	0.163	0.71 [0.61 - 0.84]	3.8×10^{-5}	IL21	Yes (Table 1)
rs12512791	4	189725728	0.702	0.414	0.354	0.77 [0.68 - 0.88]	9.2×10^{-5}	TRIML1	
rs13357969	5	150731750	0.62	0.405	0.345	0.77 [0.68 - 0.88]	9.3×10^{-5}	SLC36A2	
rs7708940	5	150769244	0.379	0.406	0.344	0.77 [0.67 - 0.87]	5.4×10^{-5}	SLC36A2	
rs6936059	6	52573410	0.786	0.453	0.521	1.32 [1.16 - 1.49]	1.8×10^{-5}	TRAM2	
rs4571541	6	92885445	0.43	0.245	0.303	1.34 [1.17 - 1.54]	2.8×10^{-5}	EPHA7	
rs4446534	6	92886209	1	0.317	0.379	1.32 [1.16 - 1.50]	3.3×10^{-5}	EPHA7	
rs10484716	6	128217415	0.148	0.224	0.174	0.73 [0.62 - 0.85]	6.8×10^{-5}	C6orf190	
rs648119	8	103207221	0.143	0.423	0.488	1.30 [1.15 - 1.47]	4.0×10^{-5}	NCALD	
rs10505604	8	134096770	0.268	0.235	0.299	1.39 [1.21 - 1.60]	3.1×10^{-6}	TG	No (Table S3B)
rs4469515	9	1329391	0.952	0.331	0.4	1.35 [1.19 - 1.53]	4.6×10^{-6}	DMRT2	No (Table S3B)
rs13301122	9	28266698	1	0.211	0.265	1.35 [1.17 - 1.56]	3.8×10^{-5}	LINGO2	
rs1064891	10	6316580	0.371	0.386	0.459	1.35 [1.19 - 1.53]	2.7×10^{-6}	PFKFB3	No (Table S3B)
rs1539234	10	6316749	0.435	0.387	0.458	1.34 [1.18 - 1.52]	4.6×10^{-6}	PFKFB3	
rs276892	11	3732947	0.958	0.519	0.452	0.76 [0.68 - 0.87]	2.2×10^{-5}	NUP98	
rs276885	11	3752982	0.596	0.518	0.452	0.77 [0.68 - 0.87]	3.0×10^{-5}	NUP98	
rs10838353	11	44744691	0.012	0.246	0.302	1.33 [1.16 - 1.53]	4.8×10^{-5}	TSPAN18	
rs7104791	11	110702068	0.934	0.198	0.249	1.35 [1.16 - 1.56]	7.9×10^{-5}	POU2AF1	
rs582465	11	111173795	0.956	0.392	0.332	0.77 [0.68 - 0.88]	9.7×10^{-5}	ALG9	
rs2325630	13	74115455	0.952	0.325	0.386	1.30 [1.15 - 1.48]	5.5×10^{-5}	KLF12	
rs7159238	14	81135391	0.746	0.433	0.504	1.33 [1.18 - 1.51]	6.2×10^{-6}	SEL1L	No (Table S3B)
rs7497057	15	32741278	0.014	0.185	0.139	0.71 [0.60 - 0.84]	8.5×10^{-5}	GJA9	
rs1353122	15	79993485	0.337	0.166	0.12	0.69 [0.57 - 0.83]	5.3×10^{-5}	RKHD3	
rs10500391	16	12868621	1	0.287	0.232	0.75 [0.65 - 0.87]	8.3×10^{-5}	FLJ11151	
rs3809983	18	54353748	0.199	0.47	0.543	1.34 [1.18 - 1.52]	3.7×10^{-6}	ALPK2	
rs3809982	18	54354054	0.099	0.472	0.543	1.33 [1.17 - 1.50]	7.0×10^{-6}	ALPK2	No (Table S3B)
rs3809973	18	54355912	0.254	0.423	0.486	1.29 [1.14 - 1.46]	6.6×10^{-5}	ALPK2	
rs3809970	18	54355971	0.624	0.425	0.486	1.28 [1.13 - 1.45]	9.2×10^{-5}	ALPK2	
rs12103986	18	54356242	0.277	0.423	0.485	1.28 [1.13 - 1.45]	7.8×10^{-5}	ALPK2	
rs7245277	18	74238260	0.58	0.078	0.117	1.56 [1.27 - 1.92]	1.9×10^{-5}	SALL3	
rs1036229	19	34672350	0.047	0.105	0.149	1.48 [1.23 - 1.78]	2.5×10^{-5}	POP4	
rs6053908	20	6034772	0.287	0.081	0.118	1.52 [1.24 - 1.87]	5.1×10^{-5}	C20orf42	
rs200755	20	15550333	0.803	0.307	0.251	0.76 [0.66 - 0.87]	7.9×10^{-5}	C20orf133	
rs3124045	X	122772520	0.397	0.28	0.216	0.71 [0.60 - 0.83]	2.9×10^{-5}	BIRC4	

b) Additional SNPs (outside 4q27 region) tested in Dutch and Irish collections

			UK genome wide scan 778 cases, 1422 controls			Irish 483 cases, 560 controls			Dutch 508 cases, 929 controls[1]		
SNP	**Chr**	**Position (bp, b36)**	**MAF Controls**	**MAF Cases**	**P**	**MAF Controls**	**MAF Cases**	**P**	**MAF Controls**	**MAF Cases**	**P**
rs1021621	1	165465160	49.1%	42.5%	3.6×10^{-5}	44.7%	45.3%	0.79	47.9%	48.4%	0.79
rs6762743	3	180494694	29.3%	22.4%	8.1×10^{-7}	28.1%	24.5%	0.075	27.0%	24.7%	0.22[1]
rs9290678	3	180499714	29.2%	22.4%	1.1×10^{-6}	28.6%	24.0%	0.021	26.0%	24.7%	0.42
rs10505604	8	134096770	23.5%	29.9%	3.1×10^{-6}	23.2%	24.2%	0.61	24.2%	24.6%	0.82[1]
rs4469515	9	1329391	33.1%	40.0%	4.6×10^{-6}	36.0%	38.7%	0.23	37.9%	37.7%	0.94[1]
rs1064891	10	6316580	38.6%	45.9%	2.7×10^{-6}	39.9%	40.5%	0.79	43.3%	44.8%	0.50[1]
rs7159238	14	81135391	43.3%	50.4%	6.2×10^{-6}	45.8%	45.9%	0.96	45.6%	44.9%	0.73[1]
rs3809982	18	54354054	47.2%	54.3%	7.0×10^{-6}	50.4%	49.6%	0.75	49.3%	53.1%	0.057

[1] For these SNPs n= 569 Dutch controls were genotyped.

MAF Minor allele frequency
HWE Hardy Weinberg equilibrium exact test
P values P values are from two-tailed chi-squared allele count tests.

Supplementary table 4

4q27 SNPs and haplotypes in three collections

a) Detailed single SNP association results

SNP, Position, Cohorts	**Platform/ Assay**	**Samples**	**Call rate**	**Genotypes (n)**			**MAF**	**P Value**
rs6835946, chromosome 4, 123289289 bp				**AA**	**AG**	**GG**		
UK celiac cases	Infinium II	778	99.60%	67	335	373	30.30%	0.63
UK population controls	Infinium II	1422	99.60%	119	599	698	29.60%	
Dutch celiac cases	Taqman	508	98.00%	38	174	286	25.10%	
Dutch healthy controls	C-29841309_10	929	94.30%	61	325	490	25.50%	0.81
Irish celiac cases	Taqman	483	97.90%	40	217	216	31.40%	
Irish healthy controls	C-29841309_10	560	98.20%	65	238	247	33.50%	0.32
rs4374642, chromosome 4, 123320561 bp				**CC**	**CT**	**TT**		
UK celiac cases	Infinium II	778	100.00%	8	147	623	10.50%	0.02
UK population controls	Infinium II	1422	100.00%	12	214	1196	8.40%	
Dutch celiac cases	Taqman	508	98.60%	11	99	391	12.10%	
Dutch healthy controls	C-30678902_10	929	98.30%	7	139	767	8.40%	0.002
Irish celiac cases	Taqman	483	99.00%	3	73	402	8.30%	
Irish healthy controls	C-30678902_10	560	99.10%	0	74	481	6.70%	0.17
rs13151961, chromosome 4, 123334952 bp				**GG**	**AG**	**AA**		
UK celiac cases	Infinium II	778	100.00%	14	168	596	12.60%	
UK population controls	Infinium II	1422	100.00%	42	424	956	17.90%	5.2×10^{-6}
Dutch celiac cases	Taqman	508	98.80%	12	98	392	12.20%	
Dutch healthy controls	C-26024001_10	929	99.40%	32	288	603	19.10%	2.2×10^{-6}
Irish celiac cases	Taqman	483	97.90%	8	124	341	14.80%	
Irish healthy controls	C-26024001_10	560	98.80%	24	167	362	19.40%	0.006

rs4505848, chromosome 4, 123351942 bp				**GG**	**AG**	**AA**		
UK celiac cases	Infinium II	778	99.90%	99	337	341	34.40%	0.15
UK population controls	Infinium II	1422	99.90%	151	615	654	32.30%	
Dutch celiac cases	Taqman	508	97.80%	66	242	189	37.60%	
Dutch healthy controls	C-34908_30	929	97.50%	106	420	380	34.90%	0.15
Irish celiac cases	Taqman	483	96.90%	36	221	211	31.30%	
Irish healthy controls	C-34908_30	560	97.30%	39	225	281	27.80%	0.084
rs13119723, chromosome 4, 123437763 bp				**GG**	**AG**	**AA**		
UK celiac cases	Infinium II	778	99.70%	13	131	632	10.10%	
UK population controls	Infinium II	1422	99.60%	42	363	1012	15.80%	$2.0x10^{-7}$
Dutch celiac cases	Taqman	508	99.00%	12	92	399	11.50%	
Dutch healthy controls	C-26404981_10	929	98.90%	32	238	649	16.40%	0.000
Irish celiac cases	Taqman	483	99.40%	10	103	367	12.80%	
Irish healthy controls	C-26404981_10	560	99.30%	23	134	399	16.20%	0.03
rs1127348, chromosome 4, 123500310 bp				**CC**	**CT**	**TT**		
UK celiac cases	Infinium II	778	100.00%	63	260	455	24.80%	0.016
UK population controls	Infinium II	1422	100.00%	66	483	873	21.60%	
Dutch celiac cases	Taqman	508	97.80%	30	195	272	25.70%	
Dutch healthy controls	C-12001860_1	929	98.30%	48	343	522	24.00%	0.34
Irish celiac cases	Taqman	483	97.70%	17	171	284	21.70%	
Irish healthy controls	C-12001860_1	560	98.00%	13	159	377	16.90%	0.005
rs11732095,chromosome 4, 123567795 bp				**GG**	**AG**	**AA**		
UK celiac cases	Infinium II	778	99.90%	12	113	652	8.80%	
UK population controls	Infinium II	1422	100.00%	16	212	1194	8.60%	0.79
Dutch celiac cases	Taqman	508	99.00%	5	90	408	9.90%	
Dutch healthy controls	C-1866308_10	929	94.90%	9	142	731	9.10%	0.45
Irish celiac cases	Taqman	483	96.90%	5	70	393	8.50%	
Irish healthy controls	C-1866308_10	560	98.20%	6	73	471	7.70%	0.5
rs10857092,chromosome 4, 123608669 bp				**AA**	**AG**	**GG**		
UK celiac cases	Infinium II	778	100.00%	3	75	700	5.20%	
UK population controls	Infinium II	1422	100.00%	9	169	1244	6.60%	0.069
Dutch celiac cases	Taqman	508	98.00%	4	70	424	7.80%	
Dutch healthy controls	C-1866315_10	929	95.30%	5	119	761	7.30%	0.6
Irish celiac cases	Taqman	483	97.50%	2	62	407	7.00%	
Irish healthy controls	C-1866315_10	560	98.20%	4	77	469	7.70%	0.53
rs6822844,chromosome 4, 123728871 bp				**TT**	**TG**	**GG**		
UK celiac cases	Infinium II	778	100.00%	14	168	596	12.60%	
UK population controls	Infinium II	1422	100.00%	42	425	955	17.90%	$4.6x10^{-6}$
Dutch celiac cases	Taqman	508	99.60%	12	101	393	12.40%	
Dutch healthy controls	C-28983601_10	929	99.50%	31	280	613	18.50%	$2.1x10^{-5}$
Irish celiac cases	Taqman	483	96.10%	8	116	340	14.20%	
Irish healthy controls	C-28983601_10	560	97.70%	21	173	353	19.70%	0.001
rs4492018,chromosome 4, 123733978 bp				**AA**	**AG**	**GG**		
UK celiac cases	Infinium II	778	100.00%	57	300	421	26.60%	
UK population controls	Infinium II	1422	100.00%	102	549	771	26.50%	0.93
Dutch celiac cases	Taqman	508	97.20%	31	162	301	22.70%	
Dutch healthy controls	C-1597477_10	929	97.80%	50	320	539	23.10%	0.8
Irish celiac cases	Taqman	483	96.70%	32	208	227	29.10%	
Irish healthy controls	C-1597477_10	560	96.40%	53	220	267	30.20%	0.6
rs975405,chromosome 4, 123740630 bp				**CC**	**CT**	**TT**		
UK celiac cases	Infinium II	778	99.70%	126	347	303	38.60%	
UK population controls	Infinium II	1422	99.90%	256	703	461	42.80%	0.007
Dutch celiac cases	Taqman	508	97.80%	93	226	178	41.50%	
Dutch healthy controls	C-8949724_10	929	94.30%	156	447	273	43.30%	0.34
Irish celiac cases	Taqman	483	96.30%	75	221	169	39.90%	
Irish healthy controls	C-8949724_10	560	97.90%	102	277	169	43.90%	0.07
rs4833837,chromosome 4, 123756413 bp				**GG**	**GA**	**AA**		
UK celiac cases	Not tested							
UK population controls	Not tested							
Dutch celiac cases	Taqman	508	98.40%	63	219	218	34.50%	
Dutch healthy controls	C-25473096_10	769	96.70%	83	334	327	33.60%	0.64
Irish celiac cases	Not tested							
Irish healthy controls	Not tested							

				TT	CT	CC		
rs1398553,chromosome 4, 123767518 bp								
UK celiac cases	Infinium II	778	99.90%	102	326	349	34.10%	
UK population controls	Infinium II	1422	100.00%	133	603	686	30.60%	0.016
Dutch celiac cases	Taqman	508	97.00%	66	215	212	35.20%	
Dutch healthy controls	C-8949749_10	929	93.60%	91	386	393	32.60%	0.18
Irish celiac cases	Taqman	483	96.50%	42	201	223	30.60%	
Irish healthy controls	C-8949749_10	560	98.20%	35	213	302	25.70%	0.015
				GG	**AG**	**AA**		
rs2893008,chromosome 4, 123772264 bp								
UK celiac cases	Infinium II	778	100.00%	8	140	630	10.00%	
UK population controls	Infinium II	1422	100.00%	11	205	1206	8.00%	0.021
Dutch celiac cases	Taqman	508	98.40%	8	94	398	11.00%	
Dutch healthy controls	C-16088771_10	929	99.70%	4	129	793	7.40%	0.001
Irish celiac cases	Taqman	483	95.20%	1	65	394	7.30%	
Irish healthy controls	C-16088771_10	560	98.60%	0	72	480	6.50%	0.5
				TT	**CT**	**CC**		
rs6840978,chromosome 4, 123774157 bp								
UK celiac cases	Infinium II	778	100.00%	23	208	547	16.30%	
UK population controls	Infinium II	1422	100.00%	61	489	872	21.50%	3.8×10^{-5}
Dutch celiac cases	Taqman	508	98.60%	11	130	360	15.20%	
Dutch healthy controls	C-1597502_10	929	94.30%	41	301	534	21.90%	2.0×10^{-5}
Irish celiac cases	Taqman	483	97.70%	19	143	310	19.20%	
Irish healthy controls	C-1597502_10	560	98.60%	33	199	320	24.00%	0.008

MAF Minor allele frequency

P Value P values are from two-tailed chi-squared allele count tests.

b) Detailed haplotype analysis results

Block 1[1]

rs6835946	rs4374642	rs13151961	rs4505848	rs13119723	rs1127348	rs11732095	rs10857092	rs6822844	UK genome scan collection Case MAF	UK genome scan collection Control MAF	UK genome scan collection P Value	Irish collection Case MAF	Irish collection Control MAF	Irish collection P Value	Dutch collection Case MAF	Dutch collection Control MAF	Dutch collection P Value
1	1	1	1	1	1	1	2	2	29.3%	29.1%	0.87	31.7%	33.5%	0.40	24.7%	24.9%	0.88
2	1	1	2	1	2	1	2	2	23.2%	20.5%	0.040	20.0%	16.0%	0.018	24.1%	22.5%	0.32
2	1	2	1	2	1	1	2	1	10.0%	15.3%	1.2×10^{-6}	11.7%	16.1%	0.0045	11.1%	15.9%	5×10^{-4}
2	2	1	1	1	1	1	2	2	10.3%	8.3%	0.027	7.8%	6.5%	0.26	11.7%	8.3%	0.0026
2	1	1	1	1	1	2	2	2	8.6%	8.2%	0.65	8.6%	7.5%	0.33	10.0%	9.0%	0.39
2	1	1	2	1	1	1	2	2	6.0%	5.2%	0.25	5.5%	4.7%	0.43	6.3%	5.6%	0.44
2	1	1	2	1	1	1	1	2	4.4%	5.2%	0.20	5.2%	6.5%	0.20	5.8%	5.7%	0.87
2	1	1	1	1	1	1	2	2	3.7%	3.6%	0.94	4.6%	4.4%	0.85	3.0%	3.2%	0.75
2	1	2	1	1	1	1	2	1	2.2%	2.2%	0.98	2.3%	3.3%	0.20	1.1%	2.6%	0.007

Block 2[1]

rs4492018	rs975405	rs1398553	rs2893008	rs6840978	UK genome scan collection Case MAF	UK genome scan collection Control MAF	UK genome scan collection P Value	Irish collection Case MAF	Irish collection Control MAF	Irish collection P Value	Dutch collection Case MAF	Dutch collection Control MAF	Dutch collection P Value
2	1	1	1	2	34.1%	30.5%	0.014	29.5%	30.5%	0.61	34.9%	32.7%	0.23
1	1	2	1	2	26.5%	26.4%	0.93	30.1%	25.5%	0.019	22.7%	23.0%	0.85
2	2	2	1	1	16.3%	21.4%	4.5×10^{-5}	19.4%	24.0%	0.012	15.1%	21.9%	1.6×10^{-5}
2	2	2	1	2	12.4%	13.4%	0.33	13.0%	13.1%	0.99	15.2%	14.0%	0.41
2	2	2	2	2	10.1%	8.0%	0.022	7.5%	6.6%	0.45	11.0%	7.5%	0.0017

[1] Block structure (determined in Haploview using data from these 14 SNPs and applying strict Gabriel *et al* criteria) was identical in UK and Irish collections. The UK / Irish structure was applied for analysis of the Dutch dataset which showed a minor difference in the position of the block 1 / block 2 boundary.

MAF Minor allele frequency

7 Genetic variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis

Nature Genetics, 2008 Jan;40(1):29-31.

Lude H. Franke[1,§], Michael A. van Es[2,§], Paul W.J. van Vught[2,§], Hylke M. Blauw[2,§], Christiaan G.J. Saris[2], Ludo van den Bosch[3], Sonja W. de Jong[2], Vianney de Jong[4], Frank Baas[5], Ruben van 't Slot[1], Robin Lemmens[3], Helenius J. Schelhaas[6], Anna Birve[7], Kristel Sleegers[8,9], Christine Van Broeckhoven[8,9], Jennifer C. Schymick[10], Bryan J. Traynor[11], John H.J. Wokke[2], Cisca Wijmenga[1,12], Wim Robberecht[3], Peter M. Andersen[7], Jan H. Veldink[2], Roel A. Ophoff[13,14], Leonard H. van den Berg[2]

1 Complex Genetics Section, Department of Biomedical Genetics and
2 Department of Neurology, Rudolf Magnus Institute of Neuroscience, University Medical CenterUtrecht, Utrecht 3584 CX, The Netherlands.
3 Department of Neurology, University Hospital Gasthuisberg, Leuven B-3000, Belgium.
4 Departments of Neurology and
5 Neurogenetics, Academic Medical Center, Amsterdam 1105 AZ, The Netherlands.
6 Department of Neurology, Radboud University Nijmegen Medical Centre, Nijmegen 6525 GA, The Netherlands.
7 Institute of Clinical Neuroscience, Umea University Hospital, Umea SE-901 85, Sweden.
8 Neurodegenerative Brain Diseases Group, Department of Molecular Genetics, VIB, Antwerpen B-2610, Belgium.
9 University of Antwerp, Antwerpen B-2610, Belgium.
10 Laboratory of Neurogenetics, National Institute of Aging, National Institutes of Health, Bethesda, Maryland 20892, USA.
11 Section on Developmental Genetic Epidemiology, National Institute of Mental Health, Bethesda, Maryland 20892, USA.
12 Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen 9700 RB, The Netherlands.
13 Department of Medical Genetics and Rudolf Magnus Institute of Neuroscience, University Medical Center Utrecht, Utrecht 3584 CX, The Netherlands.
14 Neuropsychiatric Institute, University of California, Los Angeles, California 90095, USA.
§ These authors contributed equally to this work.

Summary

We identified a SNP in the DPP6 gene that is consistently strongly associated with susceptibility to amyotrophic lateral sclerosis (ALS) in different populations of European ancestry, with an overall P value of 5.04×10^{-8} in 1,767 cases and 1,916 healthy controls and with an odds ratio of 1.30 (95% confidence interval (CI) of 1.18–1.43). Our finding is the first report of a genome-wide significant association with sporadic ALS and may be a target for future functional studies.

Amyotrophic lateral sclerosis (ALS) is a severely disabling and lethal disorder caused by progressive degeneration of motor neurons in the brain, spinal cord and brainstem. ALS affects 1-3 per 100,000 people, and the average survival time is three years from disease onset. To date, no effective treatment is available[1]. Familial ALS accounts for up to 10% of ALS cases, with approximately 20% of familial ALS cases linked to mutations in SOD1. Mutations in ALS2, DCTN1, VAPB and ANG have been found in rare cases of familial ALS[2]. Sporadic ALS accounts for 490% of ALS cases and is considered to be a multifactorial disease with an estimated heritability ranging from 0.38 to 0.85 [3]. Variants in several genes, including ANG[4], VEGF[5], HFE[6] and PON1 [7] and copy number variations in SMN1 and SMN2 [8] have been reported to be associated with ALS susceptibility. However, attempts to replicate these findings in other populations have frequently failed. For instance, sequence variations in ANG are reported to be associated with ALS in Irish and Scottish populations but rarely in English, Swedish or Italian populations. Similarly, mutations in SOD1 are found in 12% - 23% of families with ALS in the United States, the UK, Germany, Sweden and Belgium, but they are rare in families with ALS in Portugal, Switzerland and The Netherlands (F.B. and P.M.A., unpublished data). Sporadic as well as familial ALS therefore seems to be a genetically heterogeneous disease, even across European populations.

To identify previously unknown ALS susceptibility genes, we carried out a genome-wide association study using Illumina 300K Beadchips (Supplementary Methods). After stringent quality control, we carried out association analysis on 311,946 SNPs in 461 affected individuals (cases) and 450 healthy controls matched in age, gender and ethnicity from The Netherlands. The overall call rate was 99.5%. Results for all 311,946 SNPs are available online (www.alscentrum.nl/index.php?id=GWA). We did not observe any genome-wide significant association with ALS after Bonferroni correction for multiple testing. Recently, first-stage data from a genome-wide association study of 276 ALS cases and 271 controls from the United States was released. No genome-wide significant findings were observed in this study[9]. Considering the genetic heterogeneity of ALS and the fact that both studies were conducted with relatively small sample sizes, we hypothesized that signals from truly associated SNPs might be present, although weak. We therefore decided to combine both datasets and to follow up on all SNPs that had $P < 0.01$ in each study independently and unidirectional allelic association (that is, association in the same direction of the allele associated). Fifteen SNPs fulfilled these criteria and were analyzed in three additional independent populations consisting of 272 cases and 336 controls from The Netherlands, 467 cases and 437 controls from Sweden and 291 cases and 420 controls from Belgium (Supplementary Methods and Supplementary Table 1). Overall power for the study is shown in Supplementary Table 2. Genotyping of these 15 SNPs was done with Taqman technology (Supplementary Methods). We included 100 randomly selected individuals from our Dutch genome-wide association (GWA) study sample for TaqMan genotyping of the 15 SNPs and observed a concordance rate of >99.6% between platforms. Before this analysis, we examined whether population stratification was present in the available GWA data from the Dutch and US sample series using Eigenstrat and did not detect any (see Supplementary Methods and Supplementary Figure 1a,b). Because the sample series in our study were derived from different populations, we calculated overall P values and odds ratios (OR) using the Mantel-Haenszel method as well as the χ^2 test on allele counts for all 15 SNPs.

Only one SNP, rs10260404, showed genome-wide significance after Bonferroni correction for the 311,946 SNPs tested in the first stage. The overall P value for rs10260404 was 5.04×10^{-8} (corrected $P = 0.017$) with an odds ratio of 1.30 (95% CI = 1.18ñ1.43) using the χ2 test and 5.40×10^{-8} with an OR of 1.30 (95% CI = 1.18ñ1.43) using the Mantel-Haenszel method[10]. The association for rs10260404 was slightly more significant under a genotypic model (Cochran-Armitage trend test), with a P value of 3.30×10^{-8} and with an increased disease susceptibility for homozygote carriers

Table 1

Descriptive statistics and results for SNP rs10260404

	Number cases	Number controls	MAF Cases[a]	MAF Controls[a]	HWE Cases[b]
Stage 1					
Netherlands	461	450	0.44	0.37	0.48
USA	276	271	0.42	0.34	0.28
Stage I combined[e]	737	721	0.43	0.36	0.39
Stage II					
Netherlands	272	336	0.42	0.37	0.18
Sweden	467	439	0.4	0.34	0.26
Belgium	291	420	0.4	0.35	0.43
Stage II combined[e]	1,030	1,195	0.41	0.35	0.11
Stages I+II combined[e]	1,767	1,916	0.42	0.35	0.33

[a] MAF: Minor allele frequencies

[b] HWE: Hardy-Weinberg equilibrium P values

[c] P values were calculated for each individual population using χ^2 test on allele counts.

[d] Odds ratios (OR) were calculated for the minor allele in each population; 95% confidence intervals are shown in parentheses.

[e] P values and ORs using data from multiple populations were calculated using the Mantel-Haenszel method.

HWE Controls[b]	P value	OR (95%CI)[d]	
0.78	0.006	1.3	(1.08–1.56)
0.06	0.003	1.45	(1.13–1.86)
0.19	4.30×10^{-5}	1.34	(1.08–1.65)
0.51	0.04	1.26	(1.01–1.58)
0.23	0.006	1.31	(1.08–1.58)
0.39	0.11	1.21	(0.96–1.51)
0.06	0.000	1.26	(1.11–1.42)
0.38	5.40×10^{-8}	1.3	(1.18–1.43)

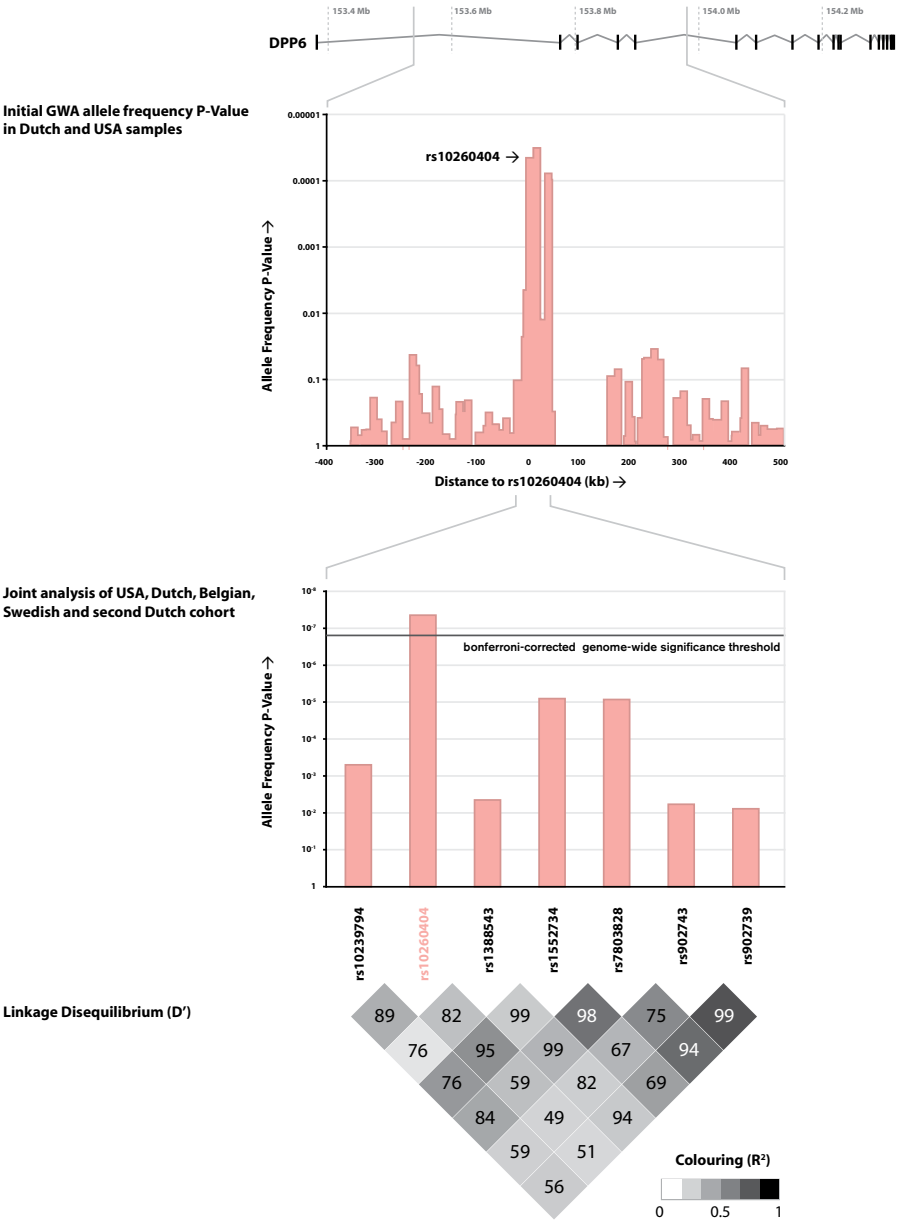
of the risk allele (OR = 1.60 with 95% CI = 1.32 - 1.92) compared to heterozygotes (OR = 1.20 with 95% CI = 1.06 - 1.41) in a dose-dependent manner. P values and odds ratios for each individual population are shown in Table 1. The minor allele frequency for rs10260404 was 42% for cases compared to 35% for controls. Results for all 15 SNPs analyzed in stage 2 are shown in Supplementary Table 3. Rs10260404 maps to a 50-kb linkage disequilibrium (LD) block on chromosome 7q36 ($r^2 = 0.8$), within a gene encoding dipeptidyl peptidase 6 (DPP6; Fig. 1).

Combining the two GWA sets, we found several SNPs within this 50-kb block that showed association with disease at P = 0.01. To rule out LD beyond this 50-kb block, we re-examined 130 SNPs in a 900-kb region surrounding rs10260404 and did not find any SNP to be associated at P = 0.01 (Fig. 1). Comparison of LD structure in this 900-kb region showed similar haplotype structure in the Dutch, US and HapMap CEPH sample datasets (Supplementary Fig. 2a). Further examination of the associated 50-kb LD block also indicated that similar LD structure is present in both the Dutch and US population (Supplementary Fig. 2b,c). It is therefore unlikely that the initial finding of ALS association is due to genetic variation outside the 50-kb LD block containing rs10260404. To fine-map the associated 50-kb LD block and carry out haplotype analyses, we additionally genotyped all SNPs (n = 6) within this block that showed an association with disease at $P < 0.01$ in the combined analysis of both genome-wide studies. Genotyping of these six SNPs was done with Taqman technology (Supplementary Methods). Single-SNP analysis of these six additionally genotyped SNPs did not show any SNP to be associated more significantly than rs10260404 (Supplementary Table 4). We then applied a recently developed multi-marker indirect association method that takes advantage of the correlation structure between SNPs in the HapMap sample (weighted haplotype analysis (WHAP); http://whap.cs.ucla.edu) using rs10260404 and the additional six flanking SNPs[11]. Using this imputation method, we again identified the strongest association signal for rs10260404, with

Figure 1

Schematic overview of DPP6

P values from the combined analysis of the two genomewide studies are shown for all SNPs in a 900-kb region surrounding rs10260404. rs7803828, located distal to rs10260404, had a lower P value in the combined analysis of the GWAs but did not fulfill the initial criteria for SNP selection (P = 0.01 in both GWAs). Subsequent analysis of seven SNPs in the associated 50-kb locus ($r^2 = 0.8$) showed the lowest allelic P value for rs10260404 at $P = 5.04 \times 10^{-8}$. The Bonferroni-corrected genome-wide significance level was set at $P = 0.05 / 311{,}946 = 1.6 \times 10^{-7}$.



$P = 6.69 \times 10^{-8}$ (Supplementary Methods and Supplementary Table 5 online). Subsequent haplotype analysis with Haploview, using the 'solid spine of LD' method to define haplotypes, showed the strongest association signal for a haplotype containing the CC alleles of flanking SNPs rs10239794 and rs10260404, with a P value of 3.01×10^{-9} and an allelic OR of 1.34 (95% CI = 1.17 - 1.54; Supplementary Fig. 3 online). Results from examining long-range LD, fine mapping (including imputation analysis) and haplotype analysis all indicated that the strongest signal for association hinges on the 'C' allele of rs10260404, suggesting that the underlying variation for disease susceptibility is at this site. Because the entire associated 50-kb LD block containing rs10260404 is located within intron 3 of *DPP6*, and there are no known genes or microRNAs nearby, we consider this to be the putative ALS-associated gene (Fig. 1). *DPP6* is located on chromosome 7q36 at location 153,380,839 - 154,315,627 (Build 35). It consists of 26 exons and is 954 kb in size (OMIM 126141). *DPP6* (also known as DPPX) encodes a dipeptidylpeptidase-like protein expressed predominantly in the brain, with very high expression in the amygdala, cingulate cortex, cerebellum and parietal lobe (http://symatlas.gnf.org/SymAtlas). This peptidase regulates the biological activity of neuropeptides by converting precursors to active forms or vice versa[12]. DPP6 binds specific voltage-gated potassium channels and alters their expression and biophysical properties. Notably, differential *DPP6* gene expression has been linked to spinal cord injury in rats[13], and *DPP6* was also identified as a nervous systemñspecific gene with accelerated evolutionary rate in the primate lineage[14].

In conclusion, we identify genetic variation in the *DPP6* gene that is highly associated with ALS susceptibility in a combined sample of 1,767 cases and 1,916 healthy control subjects from European descent. The identified SNP, rs10260404, is located within an intron of *DPP6*, and no known functional variants within the gene have been yet identified. Further study will provide insight into genetic variation at this locus, its potential effect on gene function and, ultimately, its role in disease susceptibility. Identification of a common variant within *DPP6* is an exciting first step in the genetic study of sporadic ALS, and it opens up new avenues for studying the molecular basis of this devastating disease.

Acknowledgments

We are indebted to the individuals and their families who participated in this project. This project has been generously supported by The Netherlands Organisation for Scientific Research (NWO) and the 'Prinses Beatrix Fonds' (L.H.vdB.). We would also like to thank H. Kersten and M. Kersten for their generous support (L.H.vdB.) as well as J.R. van Dijk and the Adessium foundation (L.H.vdB.), the US National Institutes of Health grants GM68875 and MH078075 (R.A.O.), the Kempe Foundation (P.M.A.), the Swedish Brain Research Foundation and Bertil Hallsten (P.M.A.), the Björklund Foundation for ALS Research (P.M.A.), the Interuniversity Attraction Pole Programme P6/43 (Belgian Science Policy Office) (W.R., L.V.D.B. and C.V.B.) and the E. von Behring Chair for Neuromuscular and Neurodegenerative Disorders (W.R.). C.V.B., W.R. and L.V.D.B. are supported by the Fund for Scientific Research Flanders (FWO-F), and K.S. holds a postdoctoral fellowship of the FWO-F. P.M.A. and A.B. are supported by the 'Swedish Brain Power Foundation'. This study used data from the SNP Database at the US National Institute of Neurological Disorders and Stroke Human Genetics Resource Center DNA and Cell Line Repository (http://ccr.coriell.org/ninds). The authors thank E. Strengman, P. Sodaar, H. Veldman, H. Yigittop, W. Scheveneels, A. D'hondt, P. Tilkin and A. Nilsson for assistance with genotyping and DNA preparation. We also thank F.G. Jennekens and G. Hille Ris Lambers for helping with the DNA sample collection.

References

1 Rowland, L.P. & Shneider, N.A. N. Engl. J. Med. 344, 1688–1700 (2001).
2 Pasinelli, P. & Brown, R.H. Nat. Rev. Neurosci. 7, 710–723 (2006).
3 Graham, A.J., Macdonald, A.M. & Hawkes, C.H. J. Neurol. Neurosurg. Psychiatry 62, 562–569 (1997).
4 Greenway, M.J. *et al.* Nat. Genet. 38, 411–413 (2006).
5 Lambrechts, D. *et al.* Nat. Genet. 34, 383–394 (2003).
6 Sutedja, N.A. *et al.* Arch. Neurol. 64, 63–67 (2007).
7 Saeed, M. *et al.* Neurology 67, 771–776 (2006).
8 Veldink, J.H. *et al.* Neurology 65, 820–825 (2005).
9 Schymick, J.C. *et al.* Lancet Neurol. 6, 322–328 (2007).
10 Skol, A.D., Scott, L.J., Abecasis, G.R. & Boehnke, M. Nat. Genet. 38, 209–213 (2006).
11 Zaitlen, N., Kang, H.M., Eskin, E. & Halperin, E. Am. J. Hum. Genet. 80, 683–691 (2007).
12 Wada, K. *et al.* Mamm. Genome 4, 234–237 (1993).
13 Tachibana, T., Noguchi, K. & Ruda, M.A. Neurosci. Lett. 327, 133–137 (2002).
14 Dorus, S. *et al.* Cell 119, 1027–1040 (2004).

Supplementary Material

Study populations & diagnosis

Supplementary Material Study populations & diagnosis. We analyzed populations from The Netherlands, USA, Belgium and Sweden. Adhering to the principles of the Declaration of Helsinki (1964), with written informed consent and approved by the local ethical committees for medical research, venous blood samples were drawn and DNA extracted according to standard procedures. All samples from Belgian and Swedish patients were screened for SOD1 gene mutations, the Swedish samples were also screened for ANG mutations. No samples with mutations in these genes were included in this study. Since no SOD1 gene mutation has ever been reported in SALS or FALS in The Netherlands, no SOD1 screening were done in these samples[1]. All patients were diagnosed with sporadic ALS according to the 1994 El Escorial criteria[2]. Patients were included in the study when they fulfilled the criteria for probable ALS or higher. All controls had negative medical and family histories for neurodegenerative disorders. Controls were matched for age, gender and ethnicity.

Dutch population: The 461 sporadic ALS cases included in the genome wide association study were individuals referred to the University Medical Center Utrecht (UMCU), the Academic Medical Center Amsterdam (AMC) or the University Medical Center Nijmegen, St Radboud. The 450 controls included in the genome wide association study were unrelated, age- and sex-matched healthy volunteers accompanying non-ALS patients to the UMCU neurology out patient clinic and spouses of sporadic ALS patients. 272 cases and 336 controls in the second, independent Dutch population were recruited from an ongoing, prospective population based study on ALS in The Netherlands. In this study, a capture-recapture design is used to identify all prevalent and incident cases in The Netherlands. Family practitioners are then asked to recruit age- and sex-matched controls from their patient registers for each case in their practice. Belgian population: 291 individuals with sporadic ALS were unrelated and from self reported Flemish descent for at least three generations. All patients were referred to the University Hospital Gasthuisberg, Leuven. The Belgian control group existed of 420 unrelated, healthy, Flemish individuals that were selected among married-in indi-

Supplementary table 1

Baseline characteristics for the studied population

Study Populations	Total	Male (%)	Spinal Onset (%)	Age at Onset (yrs)*	Survival (months)§
The Netherlands, GWA (Stage I)					
ALS	461	59	69	59 (20-86)	33 (5-147)
Healthy Controls	450	59	-	60 (22-87)	-
USA, GWA (Stage I)					
ALS	276	63	78	55 (26-87)	
Healthy Controls	271	48	-	62 (25-94)	-
The Netherlands, 2nd population (Stage II)					
ALS	272	57	71	58 (16-83)	43 (4-196)
Healthy Controls	336	57	-	59 (29-95)	-
Belgium (stage II)					
ALS	291	59	73	59 (18-86)	39 (5-177)
Healthy Controls	420	58	-	51 (18-92)	-
Sweden (Stage II)					
ALS	467	57	66	60 (20-89)	31 (9-108)
Healthy Controls	439	52	-	62 (25-94)	-
Total					
ALS	1767	58	71	59 (16-89)	36 (4-196)
Healthy Controls	1916	57	-	58 (18-95)	-

* Age at onset is shown in years with range in parenthesis.

§ Survival is shown in months with range shown in parenthesis. No survival data is available for the US cohort.

viduals in families with neurological diseases collected for genetic studies. Swedish population: The Swedish cohort consisted of 467 cases and 437 controls. Individuals with sporadic ALS were unrelated Swedish citizens who reported (northern) Swedish citizenship for at least three generations and were referred to the Umea University ALS Clinic. The Swedish control samples were spouses of the patients or unrelated healthy controls matched for age and gender. US population: Genotype data and population characteristics for the American samples were provided digitally by Bryan J Traynor. In brief, DNA from cases and controls was obtained from the NINDS Neurogenetics Repository at the Coriell Institute for Medical Research, NJ, USA. All included individuals were unique, unrelated and from white, non-Hispanic ethnicity. 16% of sALS samples were negative for SOD1 mutations, the remaining samples were not screened. Controls were also obtained from NINDS Neurogenetics Repository at the Coriell Institute for Medical Research and were sampled from many different regions across the US. All participants underwent a detailed medical history interview. None had a history of neurological disease[3,4].

Supplementary table 2

Overall statistical power

MAF	OR	Power (%) P=0.01	Power (%) P = 1.0 x 10^{-5}	Power (%) P = 1.6 x 10^{-7}
0.10	1.30	84	20	5
	1.50	100	90	65
	1.70	100	99	99
0.35	1.30	100	85	58
	1.50	100	100	100
	1.70	100	100	100

Power is shown calculated over Stage I&II, total of 1,767 cases and 1,916 controls, at different minor allele frequencies (MAF) and different odds ratios (OR) at three separate P-values. A P-value of 1.6 x10^{-7} corresponds to genome-wide significance after Bonferroni correction.

Sample collection & DNA isolation

The Netherlands: Blood samples were collected in 10 ml EDTA tubes. DNA was isolated from whole blood using the autopure DNA isolation protocol from Qiagen (Qiagen, Valencia, USA). USA: DNA from cases and controls was obtained from the NINDS Neurogenetics Repository at the Coriell Institute for Medical Research, NJ, USA. DNA was extracted from Epstein-Barr virus immortalized lymphocyte cell lines using a salting out procedure. Belgium: Genomic DNA was extracted from peripheral lymphocytes using standard procedures, on a Chemagen Magnetic Separation Module 1 platform (Chemagen AG, Baesweiler, Germany). Sweden: Genomic DNA was extracted from white blood cells from peripheral blood using the QIAamp Blood Kit (Qiagen).

Genotyping methods

For the Dutch genome wide association study we genotyped 461 sALS cases and 450 healthy controls derived from the Dutch population. Genotyping experiments were performed at the University Medical Centre Utrecht using the Illumina Infinium II HumanHap300 SNP chips (Illumina, San Diego, CA). The sALS samples from the US were assayed using the Illumina Infinium II HumanHap 550 SNP chips1. 227 controls from the US were genotyped using the Illumina Infinium II HumanHap300 SNP chips and typed additionally using the Illumina 240S chip. A further 48 controls were assayed using the Illumina Infinium II HumanHap 550 SNP chip. All experiments were carried out according to the manufacturer's protocol. In short, 750 ng of DNA per sample was whole-genome amplified, fragmented, precipitated and resuspended in the appropriate hybridization buffer. Subsequently denatured samples were hybridized on Illumina BeadChips at 48°C for a minimum of 16 hours. After hybridization, the beadchips were processed for the single base extension reaction and stained. Chips were then imaged using the Illumina Bead Array Reader. For each sample, normalized bead intensity data was loaded into Illumina Beadstudio 2.0 and converted into genotypes. Genotypes were called using the auto-calling algorithm in Illumina Beadstudio 2.0. The SNPs, which were selected for replication in additional populations, were genotyped using Taqman allelic discrimination assays. PCR was carried out with mixes consisting of 10 ng of genomic DNA, 1 x Taqman master mix (Applied Biosystems), 1x assay mix (Applied Biosystems, Foster City, USA) and ddH2O in a 5 µl reaction volume in 384-well plates (Applied Biosystems). PCR conditions were as follows: denaturation at 95°C for 10 minutes, followed by 40 cycles of denaturation at 92 °C for 15 seconds and annealing and extension at 60°C for 1 minute. Allelic PCR products were analyzed on the ABI Prism 7900HT Sequence Detection System using SDS 2.3 software (Applied Biosystems). Primer and probe sequences are available upon request. We genotyped 100 random individuals from the Dutch genome wide association study for all 15 selected SNPs using the Taqman allelic discrimination assays to ensure

Supplementary table 3

P-values for all 15 SNPs with P < 0.01 in Stage I

SNP	Chr	Gene	GWA	GWA	Stage II P-Value				MAF		Overall
			NL	USA	NL	Be	Swe	Combined	Cases	Controls	
rs10260404	7	DPP6	0.006	0.003	0.04	0.11	0.006	0.004	0.42	0.35	5.04×10^{-8}
rs3825776	15	LIPC	0.008	0.003	0.005	0.71	0.21	0.009	0.35	0.29	8.75×10^{-6}
rs7580332	2	No Gene	0.005	0.002	0.36	0.56	0.08	0.08	0.40	0.45	8.78×10^{-6}
rs973807	8	NSMAF	0.0002	0.006	0.81	0.74	0.50	0.68	0.30	0.35	0.0006
rs1061947	17	COL1A1	0.005	0.002	0.49	0.82	0.93	0.74	0.18	0.15	0.002
rs9409314	9	No Gene	0.001	0.010	0.70	0.85	0.44	0.86	0.38	0.34	0.003
rs9380343	6	No Gene	0.003	0.001	0.85	0.84	0.77	0.88	0.03	0.05	0.003
rs10438933	18	No Gene	0.003	0.003	0.20	0.35	0.21	0.93	0.14	0.12	0.004
rs5924655	23	PASD1	0.009	0.003	0.57	0.68	0.49	0.80	0.25	0.28	0.006
rs1574549	7	CALN1	0.004	0.001	0.44	0.93	0.18	0.77	0.32	0.29	0.008
rs1493282	17	ASPA	0.007	0.002	1.00	0.50	0.79	0.95	0.26	0.23	0.007
rs12861395	23	PASD1	0.007	0.003	1.00	0.31	0.41	0.28	0.25	0.28	0.03
rs895459	2	BARD1	0.010	0.003	0.11	0.67	0.44	0.68	0.37	0.39	0.07
rs1123319	23	PASD1	0.004	0.001	1.00	0.31	0.41	0.45	0.22	0.25	0.08
rs11127401	2	No Gene	0.010	0.008	0.03	0.06	0.61	0.06	0.32	0.33	0.32

NL The Netherlands

Be Belgium

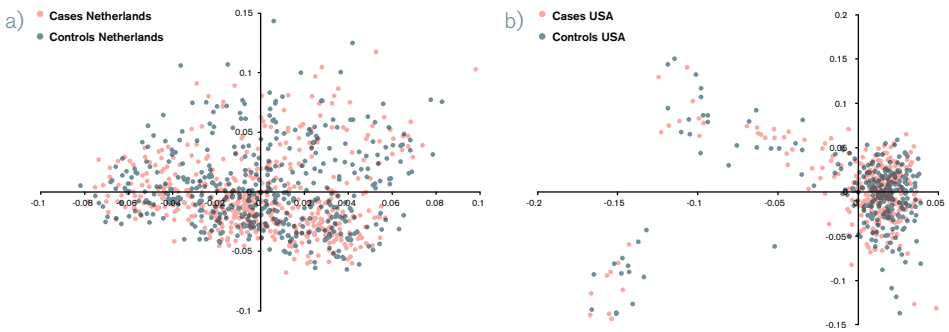
Swe Sweden

Overall MAF Minor allele frequency calculated over all five populations

OR Odds ratio, shown for each SNPs minor allele with 95% confidence interval shown in parentheses

Supplementary figure 1

a) Eigenstrat analysis between Dutch cases and controls shows no evidence for population stratification. Cases are shown in red and controls in metallic. **b)** Eigenstrat analysis between USA cases and controls shows evidence for structure, but no evidence for population stratification. Cases are shown in red and controls in metallic.



OR (9% CI)
1.30 (1.18 - 1.43)
1.34 (1.20-1.46)
0.82 (0.74-0.92)
0.82 (0.73-0.92)
1.23 (1.08-1.40)
1.16 (1.05-1.29)
0.69 (0.54-0.88)
1.24 (1.07-1.43)
0.81 (0.66-0.93)
1.15 (1.04-1.28)
1.17 (1.04-1.31)
0.85 (0.72-0.97)
0.91 (0.81-1.01)
0,83 (0,69-1,02)
0.88 (0.82-1.06)

both SNP genotyping platforms generated the same genotypes for each individual (concordance rate = 99.6%).

Quality control

Extensive quality control was done on the data from Dutch genome-wide study before performing statistical analysis. In the genome-wide study we genotyped 477 unique samples from Dutch sporadic ALS cases and 472 unique Dutch controls. We excluded 34 individuals (12 cases and 22 controls) for poor quality genotyping (call rate < 95%). Two pairs of cases (n=4) were excluded due to observed family relationships (> 200.000 concordant SNPs). Two samples were genotyped twice yielding concordance of over 99.9% for each sample. The average call rate across all samples was 99.5%. The call rate was > 99% for 298,807 SNPs and > 95% for 315,293 SNPs. HWE was calculated in controls for each SNP and < 0.05 for 14,498 SNPs and < 0.01 for 3,207 SNPs. The average minor allele frequency was 26.1%. 396 SNPs had a minor allele frequency < 0.01. After pruning SNPs according to frequency (MAF < 0.01) and genotyping (missingness > 0.05, HWE < 0.01) 311,946 SNPs remained. A total of 284,182,806 unique genotype calls were made. For the US genome-wide study quality control was performed as previously described4. In short, 276 unique samples from sporadic ALS cases and 275 unique controls were genotyped. A total of 3 controls were excluded for poor quality genotyping (call rate < 95%) and 1 individual was excluded for having an African American background. One sample was genotyped twice yielding concordance of over 99.9%. The average call rate across all samples was 99.6%. The call rate was > 99% for 514,088 SNPs and > 95% for 549,062 SNPs. HWE was calculated for each SNP and was P 0.05 for 23,657 SNPs and P 0.01 for 5,911 SNPs. The average minor allele frequency was 23.7%. A total of 302,655,011 unique genotype calls were made. Only SNPs on the Illumina 300K chips with a call rate > 95% and MAF > 0.01 were included for statistical analysis.

Association analysis

Illumina Beadstudio 2.0 was used to generate genotype final report files, which we converted into pedigree and map files using the software program Perl (http://www.perl.com). Subsequently, pedigree files were loaded into PLINK 0.99s to perform association analysis using the chi-squared test on allele counts. To calculate p-values over genotype data derived from different populations we used the Maentel-Haenszel method using PLINK 0.99s[5]. P-values for the selected 15 SNPs were also calculated under different models with PLINK 0.99s. Considering the Dutch genome-wide association study was performed using the Illumina Hap300 chip with roughly 317,000 SNPs (compared to the US GWA which included > 500,000 SNPs, only the Hap300 SNPs were included for the analyses. We further excluded SNPs from analysis if: call rate <95% in either case or control samples in the US or Dutch data sets, minor allele frequency <1% in either US or Dutch sample series or for deviation from Hardy-Weinberg equilibrium (P<0.0001 in control samples). A total of 311,946 SNPs were included for analysis. This number was also used when correcting for multiple testing using the Bonferroni method. Odds ratios with 95% CI were calculated for the minor allele of each SNP.

Power calculations

Power calculation for the genome-wide studies were performed using Genetic Power Calculator (http://pngu.mgh.harvard.edu/~purcell/gpc/) based on sample size, the average observed minor allele frequency under the assumption of a multiplicative model and a prevalence of 2 per 100,000[6]. Using these parameters the Dutch genome scan was 80% powered to detect an allelic association with P<0.01 and an odds ratio of 1.44. The US genome scan was 80% powered to detect an allelic association with P<0.01 and an odds ratio of 1.57. Power was calculated for stage I using the P-value of 0.01, since this was the cut-off value set for SNP selection. In Stage II of this study 15 SNPs were selected for further analysis. We calculated the power for replication for each of these SNPs in Stage II based on the observed MAF in Stage I and odds ratios at a P-value of 0.05 and the Bonferroni corrected value of P=0.05/15=0.0033. The size of the overall replication sample provided at least 80% power to detect for each of the 15 SNPs at a P-value of 0.05 and the corrected level of 0.0033 (Results not shown). Supplementary Table 2 shows the overall power for the study at P=0.01, 1.0×10^{-5} and 1.6×10^{-7} (which corresponds to genome-wide significance P=0.05/311,946).

Haplotype analysis

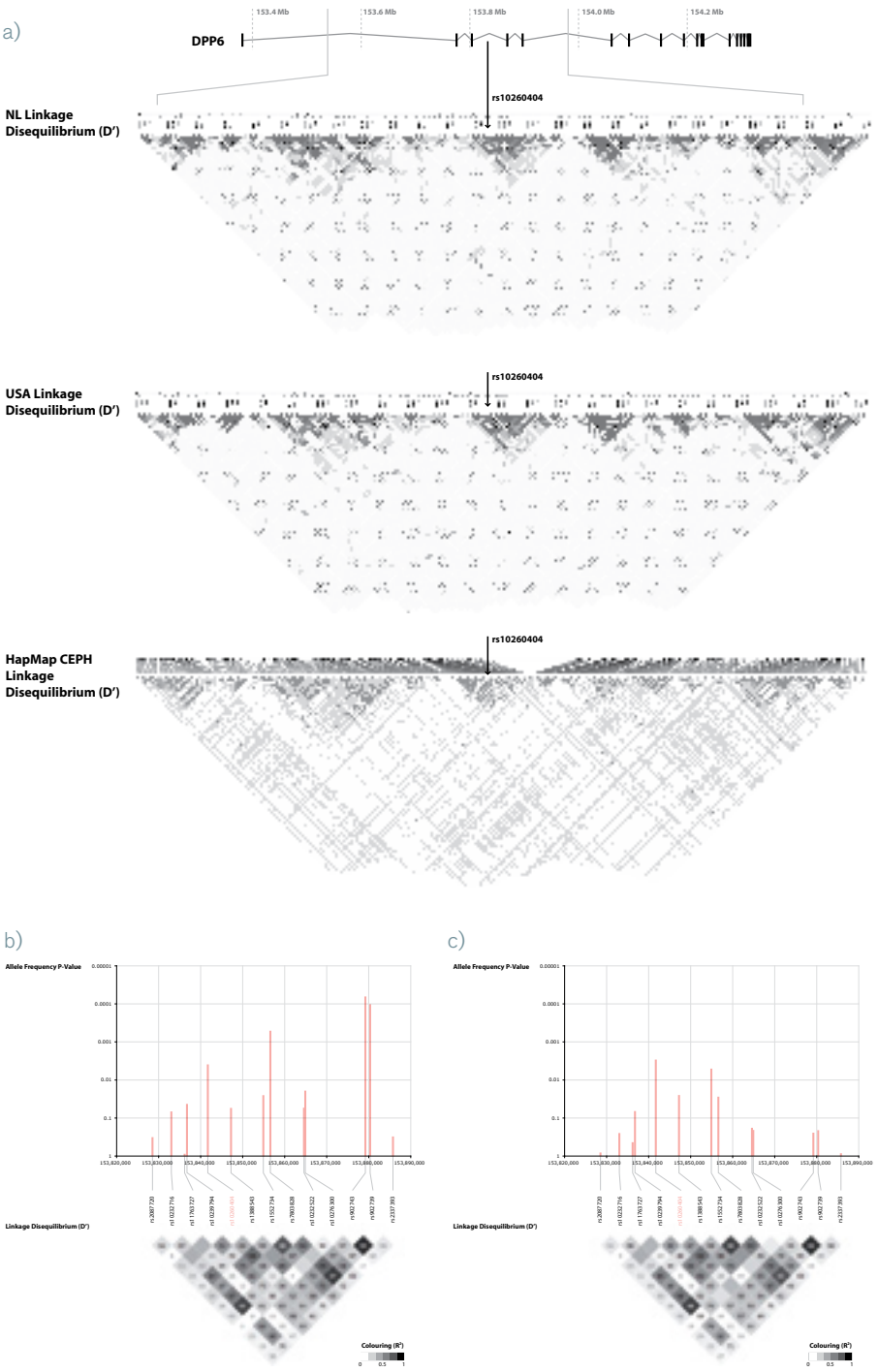
Haploview v3.32 was used for assessing linkage disequilibrium patterns and haplotype association statistics[7]. Haplotypes were defined using the solid spine of LD setting in Haploview. We included all haplotypes with a frequency>1%. P-values were calculated using a Chi-square test on alleles. Haplotypes were estimated using an accelerated EM algorithm, creating highly accurate estimations of the population frequencies of the phased haplotypes based on the maximum likelihood ratio as determined from the unphased input.

Results from population stratification analysis

We analyzed the data from both genome-wide studies for evidence of population stratification between cases and controls using Eigenstrat (Supplementary Figure 1). P-values generated by Eigenstrat were essentially identical to the initial analysis and the ranking of SNPs remained the same. Furthermore, we calculated a genomic control population inflation factor between cases and controls of 1.00. Both methods revealed no evidence for stratification between cases and controls in the genome-wide association studies. We calculated a genomic control population inflation factor between the Dutch and US population of 1.00, indicating no evidence for strong population stratification between the two populations. Furthermore,

Supplementary figure 2

a) shows a similar haplotype structure in the Dutch (NL), US and in the HapMap CEPH sample. Rs10260404 is located in a block with relatively low LD, therefore it seems unlikely that the initially observed association signal was derived from genetic variation outside of the 50 kb block of LD ($r^2 > 0.8$) in which rs102060404 is located. **b)** Top part of the figure shows P-values for SNPs in the 50 kb LD block surrounding rs10260404 in the Dutch population. The bottom part shows LD structure. **c)** Top part of the figure shows P-values for SNPs in the 50 kb LD block surrounding rs10260404 in the US population. The bottom part shows LD structure.



the case-control ratio in both studies was identical, hereby further reducing possible influence due to population stratification (461:450=1.02, cases versus controls for the Netherlands compared to 276/271=1.02, cases versus controls for the US). Even though we did not observe any evidence for population stratification, the Mantel-Haenszel method was used to calculate P-values when combining data from different populations, hereby taking into account that our overall study sample was comprised of subjects from The Netherlands, United States, Belgium and Sweden and could represent different strata.

Additional genotyping and Haplotype analysis

Considering only one SNP in the associated locus fulfilled our criteria for follow-up ($P<0.01$ in both GWAS), we hypothesized that other SNPs or haplotypes could be associated at a more significant level in the overall sample (including Stage II populations), while not necessarily most significant in the discovery set (Stage I). We therefore selected six additional tagging SNPs within the associated 50 kb locus that demonstrated an association with disease at $P<0.01$ in the combined analysis of both genome-wide studies and subsequently genotyped these SNPs in all populations. The 6 additional SNPs were rs10239794, rs1388543, rs1552734, rs7803828, rs902743 and rs902739. Individual testing of these SNPs from this region showed that none was more significantly associated than the previously identified rs10260404. However, all additionally typed SNPs consistently demonstrated P-values <0.05. Results are shown in Supplementary Table 4. Using the solid spine setting in Haploview, 2 haplotype blocks were identified. Subsequent association analysis demonstrated a P-value of 3.01×10^{-9} for the CC-haplotype of rs10260404 and rs10239794 (Supplementary Figure 3).

Weighted Haplotype analysis (WHAP)

WHAP is a statistical test for case/control association studies. It takes advantage of the known correlation structure between SNPs in the HapMap sample (www.hapmap.org) in order to improve power over traditional methods[8]. The full details of the rationale and the methodology are described in Zaitlen *et al*[9]. The WHAP tool is available online at http://whap.ucla.edu/. The analysis was performed using genotype data from all included subjects from all populations for the following SNPs: rs10239794, rs10260404, rs1388543, rs1552734, rs7803828, rs902743 and rs902739. Quality control parameters were set at 0.001 for the Hardy-Weinberg Equilibrium P-value and 10% for missing genotype or missing individual frequencies. HapMap correlations were taken from the European-American (CEPH) HapMap sample. WHAP analysis yielded the most significant association with the collected marker rs10260404 at $P=6.69 \times 10^{-8}$ (Supplementary Table 5). The most significant result from the WHAP analysis is the same SNP identified by the initial scan and fine mapping, suggesting that the true association signal represented with these seven SNPs and the imputated haplotype structure is located at very close proximity to this marker or rs10260404 itself.

References supplementary material

1 Baas, F. & Andersen, P.M. (2007).
2 Brooks, B.R. J. Neurol. Sci. 124 Suppl, 96-107 (1994).
3 Fung, H.C. *et al*. Lancet Neurol. 5, 911-916 (2006).
4 Schymick, J.C. *et al*. Lancet Neurol. 6, 322-328 (2007).
5 Purcell S., *et al*. Am. J. Hum.Gen. 81, 559-75 (2007).
6 Purcell, S., Cherny, S.S., & Sham, P.C. Bioinformatics. 19, 149-150 (2003).
7 Barrett, J.C., Fry, B., Maller, J., & Daly, M.J. Bioinformatics. 21, 263-265 (2005).
8 The International HapMap Consortium. Nature 426, 789-796 (2003).
9 Zaitlen, N., Kang, H.M., Eskin, E., & Halperin, E. Am. J. Hum. Genet. 80, 683¬691 (2007).

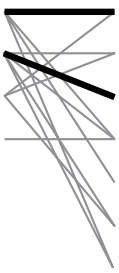
Supplementary figure 3

Using the Solid Spine of LD method in Haploview two haplotype blocks were defined. The CC-haplotype composed of rs10239794 and rs10260404 demonstrated the most significant result.

Block 1

rs10239794 rs10260404	Case Freq	Control Freq	P value
TT	0.385	0.421	0.0018
CC	0.401	0.334	3.01×10^{-9}
CT	0.199	0.226	0.0056
TC	0.015	0.02	0.1514

D' 0.69



Block 2

rs1388543 rs1552734 rs7803828 rs902743 rs902739	Case Freq	Control Freq	P value
TTTCT	0.384	0.421	0.0013
TCGTC	0.202	0.18	0.0221
CCGTC	0.197	0.169	0.0021
TCTCT	0.092	0.089	0.7085
TTTTC	0.059	0.074	0.0087
TCGCC	0.032	0.03	0.6396
CCGCC	0.015	0.017	0.5209

Deletion

100,000,000 bp

Supplementary table 4

Results for additionally genotyped SNPs

SNP	Stage I			Stage II				Overall
	NL	USA	Meta-analysis	NL	Swe	Be	Meta-analysis	P-value
rs10239794	0.045	0.039	0.004	0.29	0.06	0.54	0.06	6.10×10^{-4}
rs1388543	0.053	0.02	0.003	0.24	0.63	0.3	0.43	0.009
rs1552734	0.034	0.003	5.0×10^{-4}	0.15	0.11	0.11	0.005	7.44×10^{-6}
rs7803828	5.53×10^{-4}	0.019	3.0×10^{-5}	0.31	0.23	0.42	0.31	3.08×10^{-5}
rs902743	5.90×10^{-5}	0.196	7.5×10^{-5}	0.37	0.4	0.84	0.99	0.01
rs902739	1.01×10^{-4}	0.173	9.0×10^{-5}	0.63	0.52	0.8	0.87	0.01

NL The Netherlands
Be Belgium
Swe Sweden

P-values were calculated using the χ^2 test on allele counts. Overall P-values were calculated using the Mantel-Haenszel Method.

Supplementary table 5

Results from WHAP analysis

db SNP ID	Chr	Position	MAF	r^2	P-value	Tagset
rs10260404	7	153841730	0.38	1.00	$6.69x10^{-8}$	Collected marker
rs10264387	7	153851755	0.50	0.94	$3.75x^{10-7}$	rs10260404, rs1552734, rs902743
rs12668377	7	153840791	0.41	0.97	$1.03x^{10-6}$	rs10260404, rs1388543
rs10247061	7	153837975	0.43	0.85	$5.69x^{10-6}$	rs10239794, rs10260404, rs902739
rs11975187	7	153863909	0.39	1.00	$6.84x^{10-6}$	rs10260404, rs7803828, rs902743
rs1907616	7	153859314	0.40	0.97	$7.15x^{10-6}$	rs10260404, rs7803828, rs902743
rs6956780	7	153850432	0.21	0.68	$7.43x^{10-6}$	rs10260404, rs1388543, rs1552734
rs7803828	7	153856587	0.43	1.00	$1.19x^{10-5}$	Collected marker
rs1552734	7	153854973	0.46	1.00	$1.51x^{10-5}$	Collected marker
rs923711	7	153852675	0.45	0.97	$1.56x^{10-5}$	rs1388543, rs1552734, rs902743
rs10952491	7	153848888	0.50	0.86	$3.43x^{10-5}$	rs10260404, rs7803828, rs902743
rs6964455	7	153839596	0.48	0.83	$3.63x^{10-5}$	rs10239794, rs10260404, rs1552734
rs13246178	7	153848111	0.18	0.85	$3.90x^{10-5}$	rs10260404, rs1388543, rs1552734
rs12538549	7	153840163	0.18	0.75	$4.86x^{10-5}$	rs10260404, rs1388543, rs1552734
rs7780519	7	153848685	0.18	0.85	$5.18x^{10-5}$	rs10260404, rs1388543, rs1552734
rs13232525	7	153839757	0.18	0.71	$6.16x^{10-5}$	rs10260404, rs1388543, rs1552734
rs11767475	7	153839973	0.18	0.75	$6.58x^{10-5}$	rs10260404, rs1388543, rs1552734
rs6464429	7	153844747	0.43	0.90	$9.61x^{10-5}$	rs10239794, rs10260404, rs902743
rs12534820	7	153834965	0.16	0.70	$2.02x^{10-4}$	rs10239794, rs10260404, rs1388543
rs4725547	7	153834387	0.43	0.86	$2.20x^{10-4}$	rs10239794, rs1388543, rs902743
rs10267199	7	153834908	0.40	0.89	$2.42x^{10-4}$	rs10239794, rs7803828
rs1388540	7	153853562	0.39	0.76	$2.55x^{10-4}$	rs10260404, rs7803828, rs902743
rs10276300	7	153864766	0.10	0.75	$2.58x^{10-4}$	rs10260404, rs1388543, rs902743
rs10262182	7	153833200	0.45	0.87	$4.45x^{10-4}$	rs10239794, rs902743, rs902739
rs10952485	7	153829535	0.43	0.93	$6.3\ x^{10-4}$	rs10239794, rs10260404, rs1552734
rs10239794	7	153836826	0.42	1.00	$6.89x^{10-4}$	Collected marker
rs11976788	7	153829985	0.49	0.97	$7.59x^{10-4}$	rs10239794, rs902743, rs902739
rs6955044	7	153832247	0.50	0.94	$1.10x^{10-3}$	rs10239794, rs902743, rs902739
rs9690048	7	153852822	0.08	0.90	$1.31x^{10-3}$	rs10260404, rs1388543, rs902743
rs7809068	7	153859565	0.08	0.90	$1.31x^{10-3}$	rs10260404, rs1388543, rs902743
rs11971700	7	153864956	0.08	0.90	$1.31x^{10-3}$	rs10260404, rs1388543, rs902743
rs13228346	7	153866129	0.08	0.90	$1.31x^{10-3}$	rs10260404, rs1388543, rs902743
rs12703363	7	153846598	0.08	0.90	$1.31x^{10-3}$	rs10260404, rs1388543, rs902743
rs13226031	7	153832786	0.26	0.76	$1.58x^{10-3}$	rs10239794, rs10260404, rs1388543
rs10260955	7	153838166	0.34	0.82	$2.41x^{10-3}$	rs10260404, rs1388543, rs7803828
rs10274497	7	153869441	0.20	0.87	$2.73x^{10-3}$	rs10239794, rs1388543, rs7803828
rs923710	7	153852787	0.19	0.83	$2.96x^{10-3}$	rs10239794, rs1388543, rs7803828
rs10231561	7	153837944	0.18	1.00	$4.02x^{10-3}$	rs10239794, rs1388543
rs10464419	7	153856052	0.30	0.77	$4.15x^{10-3}$	rs10239794, rs1552734, rs902743
rs6969351	7	153877668	0.46	0.97	$4.27x^{10-3}$	rs10260404, rs902743, rs902739
rs12533032	7	153834856	0.28	0.74	$4.60x^{10-3}$	rs10239794, rs10260404, rs1388543
rs11766937	7	153836441	0.28	0.74	$4.6x^{10-3}$	rs10239794, rs10260404, rs1388543
rs1120724	7	153837511	0.08	0.63	$4.75x^{10-3}$	rs10239794, rs1552734, rs902739
rs1388543	7	153847278	0.204	1.00	$6.29x^{10-3}$	Collected marker
rs12703360	7	153831979	0.292	0.74	$6.55x^{10-3}$	rs10239794, rs10260404, rs1388543
rs6976924	7	153850457	0.125	1.00	$6.85x^{10-3}$	rs10239794, rs10260404, rs1552734

MAF Minor allele frequency

Results from WHAP analysis showing the top 50 SNP locations; All non collected markers are imputed based on haplotype correlations present in the European-American CEPH HapMap sample.

8 Discussion

Introduction

In this thesis we have described two genome-wide association analyses performed to identify new susceptibility genes in celiac disease[1] and amyotrophic lateral sclerosis[2]. We have also described new statistical methods that can help to identify such genes[3]. This final chapter discusses the key events in genetics that eventually enabled us to identify these disease genes and describes what types of future studies might help us to discover more. We outline how new high-throughput methodologies might contribute to further the identification and biological understanding of these susceptibility genes. Finally, some pressing ethical issues are discussed, now that commercial entities are starting to capitalize on our genome-wide findings and those of others.

Retrospective view

The history of the search for disease genes goes back to the beginning of the 20th century. In 1903 Sutton[4] and de Vries[5], and in 1904 Boveri[6], suggested that genetic variation could be exchanged between homologous chromosomes during meiosis (homologous recombination). This led Thomas Hunt Morgan in 1911 to the concept of crossing over and genetic linkage[7; 8]. In 1937, Haldane and Bell[9] suggested that markers could help to identify human disease genes. They stated if "an equally close linkage" (as between the genes for hemophilia and color-blindness) "were found between the genes for blood groups" and that "determining Huntington's chorea, we should be able, in many cases, to predict which children of an affected person would develop this disease and to advise on the desirability or otherwise of their marriage".

Essential for identifying these disease genes through linkage analysis was the availability of markers that reflect polymorphic loci in the physical vicinity of the disease genes. Initially these markers were based on clear phenotypes (such as blood groups) and serum proteins, but molecular avenues were opened in 1953 when Watson and Crick resolved the three-dimensional helix structure of DNA[10]. (Owing to the extreme elegance of this structure, they only needed a single page in Nature to describe it). However, it took until 1978 when Kan and Dozy identified typable genetic markers that were widely present throughout the human genome[11]. In 1980 Botstein *et al* observed that restriction-fragment-length polymorphisms (RFLPs), spanning the entire genome, could efficiently be used for linkage analysis[12]. Using these markers, in 1983 Gusella *et al* were the first to identify a disease gene through linkage analysis[13]: They discovered that the gene for Huntington disease mapped to chromosome 4. Various technological improvements and novel statistical methods improved the efficacy of linkage analysis. Initial single marker linkage algorithms, developed by Elston and Stewart[14] in 1971 and Ott[15] in 1974 were complemented by more powerful statistical methods that used multiple adjacent markers[16; 17]. In 1984 Jeffreys *et al* invented DNA fingerprinting[18; 19] by observing that a certain tandem-repeat was highly polymorphic and present at many sites throughout the genome. Mullis *et al* in 1985 invented polymerase chain reaction (PCR) which allowed for amplifying DNA easily[20]. Using this PCR technique, it became possible in 1989 to easily genotype common short-tandem-repeat markers, or microsatellites, that spanned the genome (Litt and Luty[21]; Weber and May[22]).

These markers enabled the widespread use of linkage analysis and its success was remarkable: the responsible loci for many Mendelian diseases could be identified. Subsequent positional cloning of the disease genes through PCR-based sequencing technologies provided insights into the different types of mutations. Deletions, duplications, inversions, short repeats or single nucleotide polymorphisms (SNPs) were shown to cause these rare disorders[23; 24].

Due to the successes for Mendelian disorders, efforts were undertaken to identify susceptibility loci for more common but complex polygenic diseases. In 1990 Hall *et al* were the first to be successful as they discovered that a locus on 17q21 was linked to early-onset familial breast cancer[25]. However, it soon turned out to be difficult to unequivocally identify loci linked to other common disorders. Various hypotheses were presented to explain this lack of success. As linkage analysis assumed certain models of inheritance and penetrance, one explanation was that these parameters had been incorrectly estimated. Non-parametric methods (such as identity-by-descent sharing methods) were developed to overcome this, but these also proved mostly ineffective. The common variant-common disease (CD-CV) hypothesis[26] became popular and predicted that there were multiple disease susceptibility loci that each contained only a limited number of common disease susceptibility alleles. As these individual alleles were common and only conferred limited susceptibility, the power had been insufficient to detect them through linkage. This hypothesis was supported by the identification of a protective allele in APOE4 in Alzheimer's disease[27], a protective Factor V allele in deep-venous thrombosis[28] and a protec-

Table 1: Overview of recent genome-wide association papers

Indicated are papers in Nature, Nature Genetics, Science or the New England Journal of Medicine (NEJM) that appeared within one year after Duerr *et al* described the first genome-wide association study with high coverage.

Authors	**Year**	**Journal**	**Disease**	**Indentified Susceptibility Loci**
Duerr *et al*	2006	Science	Inflammatory Bowel Disease	IL23R
Hampe *et al*	2006	Nature Genetics	Crohn's Disease	ATG16L1
Sladek *et al*	2007	Nature	Type 2 Diabetes	SLC30A8, IDE/KIF11/HHEX, EXT/ALX4
Gudmundsson *et al*	2007	Nature Genetics	Prostate Cancer	AW183883/AF268618
Yeager *et al*	2007	Nature Genetics	Prostate Cancer	POU5F1P1/DG8S737
Rioux *et al*	2007	Nature Genetics	Crohn's Disease	PHOX2B/NCF4, FAM92B
Steinthorsdottir *et al*	2007	Nature Genetics	Type 2 Diabetes	CDKAL1
Scott *et al*	2007	Science	Type 2 Diabetes	CDKAL1, IGF2BP2, CDKN2A/CDKN2B
Helgadottir *et al*	2007	Science	Myocardial Infarction	CDKN2A/CDKN2B
Easton *et al*	2007	Nature	Breast Cancer	Multiple
Hunter *et al*	2007	Nature Genetics	Breast Cancer	FGFR2
Stacey *et al*	2007	Nature Genetics	Breast Cancer	Multiple
Saxena et al	2007	Science	Type 2 Diabetes	CDKN2A/CDKN2B, IGF2BP2, CDKAL1
Wellcome Trust Case Control Consortium	2007	Nature	Bipolar disorder, Coronary artery disease, Crohn's disease, Hypertension, Rheumatoid arthritis, Type 1 diabetes, Type 2 diabetes	Multiple
McPherson *et al*	2007	Science	Coronary Heart Disease	CDKN2A/CDKN2B
van Heel *et al*	2007	Nature Genetics	Celiac Disease	Il2/Il21
Moffatt *et al*	2007	Nature	Childhood Asthma	ORMDL3
Winkelmann *et al*	2007	Nature Genetics	Restless Legs Syndrome	MEIS1,BTBD9, MAP2K5/LBXCOR1
Hafler *et al*	2007	NEJM	Multiple Sclerosis	IL2RA,IL7RA

tive deletion allele in CKR5 in human immunodeficiency virus[29].

To systematically assess the CD-CV hypothesis, in 1996 Risch and Merikangas proposed that genome-wide association analysis using unrelated cases and controls could be efficiently performed[30]. However, their proposal relied heavily on the availability of a physical and linkage disequilibrium map of the human genome. The Human Genome Project was initiated in 1990 to develop this physical map and on 26th June 2001 Francis Collins, Craig Venter, Bill Clinton and Tony Blair announced a major milestone: a rough draft of the human genome sequence had been generated. Bill Clinton called it "one of the most important, most wondrous maps ever produced by humankind". The laborious task of positionally cloning identified genes could now be performed in silico. It also initiated the development of many algorithms for predicting the biological functions of certain genomic loci.

But as only a consensus DNA genome assembly had been generated[31; 32] initiatives were also required to explore genetic variation in a systematic way. As most of this variation could be attributed to SNPs that can also tag other genetic variants, plans were outlined in early 2003 to generate a linkage disequilibrium map in four different populations. The International Human Haplotype Mapping Project (HapMap) aimed to identify most of the common genetic variation using SNPs that could be easily typed by then. By systematic analysis of 270 samples, the characteristics of over 3.1 million polymorphic loci were eventually determined[33; 34].

These findings allowed two groups early on to identify susceptibility genes by analyzing thousands of these SNPs. Ozaki *et al*[35] identified lymphotoxin-alpha (LTA) as a susceptibility gene in myocardial infarction using 65,671 successfully genotyped SNPs. Two years later Klein *et al* presented results for age-related macular degeneration[36]. They successfully genotyped 105,980 SNPs and used HapMap to identify complement factor H (CFH) as a strong susceptibility gene.

These initial papers were surprisingly successful in identifying susceptibility genes and provided strong support for the CV-CD hypothesis, despite the fact that their genetic coverage was limited. (Using the SNPs on the array from Klein *et al*, a two-marker analysis could tag only 39% of HapMap Phase II SNPs ($R^2 \geq 0.8$) with a MAF above 5%)[37]. The HapMap project allowed for determining a limited set of SNPs that provided much greater genetic coverage, as it was observed that many of the human SNPs were usually in strong linkage disequilibrium with each other. Through careful selection of only 300,000 SNPs (tagging SNPs), approximately 86% of the genetic variation of the Caucasian human genome could be captured (two-marker SNP analysis tags (with an $R^2 \geq 0.8$) 86% of HapMap Phase II SNPs with a MAF above 5%)[37; 38]. This spurred the development of affordable chips containing these SNPs. Affymetrix announced its Human Mapping 500K Array Set in September 2005 and Illumina soon followed with the release of its Infinium HumanHap300 Genotyping BeadChip in January 2006. Genome-wide association was no longer solely feasible, but had become financially viable as well.

Using these chips many research groups initiated genome-wide studies, leading to the first paper utilizing these chips in October 2006[39]. Duer *et al* identified IL23R as a susceptibility gene in inflammatory bowel disease. Twelve months later 18 more genome-wide studies had identified susceptibility loci for various diseases in papers published in Nature, Nature Genetics, Science or the New England Journal of Medicine (see table 1). The strict guidelines that had been imposed by these journals[40] (see box 1) ensured that most of these findings could be replicated. As such, in December 2007 Science concluded that genetic variation was the scientific breakthrough of the year.

However, it was realized that the SNPs typed in the HapMap project could not capture all genetic variation. In 2006 Redon *et al*[41] showed that structural variants, such as deletion and duplication loci, were very common throughout the genome and that they accounted for a considerable amount of genetic variation. In approximately the

Box 1: Guidelines for replicating genetic associations

In 2007 the NCI-NHGRI Working Group on Replication in Association Studies established an extensive list of points to consider when performing genome-wide association analyses40. The following categories were defined:

- **Study information:** Design, phenotyping and DNA collection should be well-described.
- **Data issues:** Genotype data or extensive summary statistics should be made available.
- **Genotyping and quality control procedures:** Numerous genotyping and quality procedures (such as population stratification analyses) should be undertaken to ensure that identified associations are real.
- **Results:** The analysis of the data should be well described, substantiating the relevance of the reported associations.
- **Replication studies:** Replication cohorts should be well described. Multiple-testing issues should be discussed.
- **Genotyping deposition in standard databases:** If genotype data is deposited in public databases, the methodologies employed for transferral of raw data should be described.
- **Points for reviewers and authors to consider regarding priority for publication:** Various aspects, such as the importance of a finding, the sample size and statistical analyses, should be considered in order to determine the relevance of the findings.

same period, various copy number variants (CNVs) turned out be associated with AIDS[42], glomerulonephritis[43] and Crohn's disease[44]. However, most of these CNVs had not been described before and therefore had not been efficiently tagged by SNPs in HapMap. As such, the available oligonucleotide arrays were biased against these variants. To resolve this, efforts are currently underway to detect these CNVs with high sensitivity, which will allow for the design of new arrays that will eventually have a much greater genetic coverage.

From a functional perspective it is tempting to assume the CNV disease associations elicit gene-dosage dependent effects. Although other biological explanations are also likely (such as abrogation of functional products near breakpoints), clear biological implications are often lacking for associated SNPs. Some of these SNPs are non-synonymous, affect splicing or lead to preliminary stop codons, and thus result in different proteins, which immediately lead to biologically contestable hypotheses. However, for the majority of associated SNPs, these scenarios do not seem to apply. Additionally, as most associated SNPs tag common haplotypes, it is quite possible that the true disease-associated variant is not the associated tagging SNP, which complicates deciphering of the biological consequences considerably.

To improve our biological insight, different high-throughput strategies have been presented and various high-throughput sequencing techniques now allow us to resequence many susceptibility loci. This was supported in 2007 by the sequencing of two different diploid genomes[45], of which one used a new, low-cost, high-throughput technology[46]. As such this allows in-depth analyses of the associated haplotypes and is likely to reveal new variants for which functional consequences might be detected.

To gain systematic insight into the potential functional effects of genetic variants, a new strategy was proposed by Jansen and Nap in 2001[47]: genetical genomics correlates genotypes with gene expression levels in a high-throughput way. Results in maize[48], yeast[49-51], rats[52; 53], mice[48; 54; 55] and humans[56-61] provided evidence that genetic variation often affects eukaryotic gene expression levels. In 2007 Stranger *et al*[62] reported that many SNPs and CNVs, which are present in the HapMap samples, control gene expression. Keurentjes *et al*[63; 64] observed that many protein levels were also under tight genetic control. These fundamental studies have already provided evidence that certain recently identified susceptibility loci have functional expression consequences[65; 66].

Putting our Studies in Perspective

The work outlined in this thesis would have been impossible without these technological and biological advances in genetics. We are indebted to several large collaborative projects that enabled us to perform the studies described here.

We performed two genome-wide association studies using the Illumina HumanHap300 BeadChips and developed a new genotype-calling algorithm; the results led to identification of two susceptibility loci: *IL2*/*IL21* in celiac disease[1] and *DPP6* in sporadic ALS[2]. Since the discovery of *IL2*/*IL21*, associations with other auto-immune diseases have also been observed for this locus[67; 68]. Apart from IL2 and IL21, other interleukins and chemokines have been shown to affect auto-immunity and inflammatory disease susceptibility. By increasing our celiac disease sample size from 2,200 to 7,049 samples, we recently identified seven more loci. *CCR1*, *CCR2*, *CCRL2*, *CCR3*, *CCR5*, *CCXCR1*, *IL18RAP*, *IL18R1* and *IL12A* are among the interleukins and chemokines that map within these loci66. These indicate that a considerable amount of the genetic risk in celiac disease is conferred by genetic variation in genes involved in adaptive immunity. The involvement of *DPP6* (rs10260404) in sporadic ALS has recently been validated in an independent Irish cohort[69] (rs10260404, P value = 0.029, identical direction of associated allele). As most sporadic ALS patients die within three years, (there is no cure available yet), our finding may help to identify pathways that are affected by alterations within this gene for follow-up research.

However, both our studies also indicate that a considerable proportion of the heritability still remains to be explained. This suggests that not only have we had insufficient statistical power to detect common variants with smaller effects, but also that it is quite likely that rare variants are present which cannot be easily detected by genome-wide association analysis.

Inspired by the ramifications of widespread copy number variation, we conducted two studies on structural variation. We identified 1,880 SNPs with an untyped allele, and have estimated approximately 700 of these reflect deletions. Although we used Illumina HumanHap300 and HumanHap550 BeadChip arrays that are biased against CNVs, we were able to identify many new deletions by employing a new calling algorithm we have developed. Our methodology has the highest attainable resolution and indicates that most of the structural variants in the genome (especially smaller ones) still remain to be identified. New oligonucleotide arrays that cover more SNPs and are less biased against CNVs will result in the identification of many more structural variants. Additionally, we have described how our algorithm can be extended to also capture duplications and rare structural variants.

In another study we investigated recurrent cytogenetic aberrations in autistic individuals, as recent papers have provided evidence that structural variation is important in the etiology of autism[70; 71]. However, it is unclear so far how this structural variation eventually results in autism. We assumed that autism might partly reflect a contiguous gene syndrome, affecting the function of multiple consecutive genes that individually cause rather atypical symptoms, but jointly contribute to an autistic phenotype. To assess this we first determined what syndromes are known to be caused by mutations within these loci and then we identified various clinical symptoms (such as seizures and craniofacial abnormalities) that are caused by mutations in the aberrant loci more often than would be expected by chance. We concluded these symptoms are likely to co-occur with autism and that they might serve to help define genetically more homogenous (patient) groups. We expect this will help increase the power to identify new susceptibility loci.

To gain insight into the functional consequences of genetic variation we have described two different strategies to bridge genetic variation with functional biology. In our Prioritizer method we assume that the genes contributing to a specific disorder are likely to confer a function that is similar3. To assess this we reconstructed a functional human gene network and then developed a method that can prioritize positional candidate genes within susceptibility loci, identified using linkage analysis. Using this approach we proposed *NPY2R* as a plausible candidate gene in Type II diabetes[72], of which SNPs within the promoter ($P < 0.009$) were subsequently indeed shown to be associated[73]. We expect that new datasets and statistical frameworks, incorporating these datasets, will help to construct more accurate networks that concentrate on specific cell types, improving insight in the functional consequences of disease genes. We also performed a genetical genomics study in 110 celiac disease samples for which both genome-wide genotype and gene expression data was available. Through this analysis we found many genes whose expression is under strong genetic control and it allowed us to identify *IL18RAP* as the most plausible positional candidate gene in a susceptibility locus we have identified as significantly associated with celiac disease[66]. Additionally we have provided preliminary evidence that these integrative approaches will also help to identify new biological pathways.

Future Perspectives

Identifying additional susceptibility loci

The susceptibility loci we identified in sporadic ALS and celiac disease provide strong support for the CD-CV hypothesis. However, the results also indicate that the known loci in these disorders usually exert only modest effects (odds ratios between 1.2 and 1.4 when excluding the HLA in celiac disease), and explain less than 10% of the heritability. Follow-up research is thus warranted to identify the real causal variants that might exert stronger effects than the

tagging SNPs we have now identified. Additionally it can be assumed additional loci play a role in these disease.

One obvious strategy to detect common variants with even smaller effects is to increase the sample size. This suggests that collaborations between various research groups and physicians will become even more important. However, physicians might be reluctant to spend their time collecting samples from individual patients when they realize their contribution will constitute only a small proportion of the total number of samples needed for performing these large studies. The physician's importance however is likely to become even greater: for many of the identified loci it is important that accurate risks can be estimated, such that these risks can eventually be used for prognosis. To determine these, prospective follow-up studies conducted in very close collaboration with physicians will be necessary. Additionally, these studies might shed some light on subtle genotype-phenotype associations.

The power to detect susceptibility loci can also be increased by using chips that have greater genetic coverage, since it has been established that not all genetic variation is currently captured by the Illumina HumanHap330 BeadChip arrays we used[1; 2]. We expect that new arrays with better coverage will help to identify more susceptibility loci and recently developed CNV calling algorithms are likely to help as well[74; 75]. We developed a new, high-resolution, calling algorithm that can detect very small deletions. We expect high-resolution calling algorithms for duplications and methods that can assess de novo CNVs will also help to identify more loci. New statistical frameworks for imputing genotypes[76] and assessing epistasis[77] are likely to contribute even more, given that sufficient numbers of samples are will become available.

A different explanation for the limited contribution of the known susceptibility loci is that a considerable proportion of the currently unknown heritability is not due to common variants, but to rare ones. So far the costs of the available technologies have been too high to assess this systematically. With the arrival of high-throughput sequencing methods[78], it is now becoming feasible to conduct these types of studies, as has been recently shown for various rare variants contributing to quantitative metabolic traits[79-82]. The currently available, high-throughput, sequencing technologies can generate millions of sequence reads. This is a major improvement (both technically and financially) over chain terminator sequencing (Sanger sequencing). However, the limited read size (currently generally up to ~250 bp) restricts analysis to approximately 2% of the human genome at a time. This requires the input DNA to be limited to loci that are deemed interesting. Common strategies involve the selection of a limited number of candidate loci or only exons[83-86]. Another limitation is that these methods can only sequence a limited number of different individuals at the same time. To overcome this, methods have recently been introduced that can simultaneously assess many samples through nucleotide bar-coding designs[87-90].

It is expected that these technologies will help to start discovering rare variants. These studies might also help to explain why there has been such a strong discrepancy between the susceptibility loci that have been identified through linkage analysis and the loci that are now being discovered through genome-wide association studies. Contrary to association analysis, linkage analysis for a given locus does not require that the susceptibility alleles (which potentially can be multiple rare variants) reside on a limited number of haplotypes. High-throughput sequencing might help to uncover what is going on in these loci and unveil new genes involved in disease.

Susceptibility loci and their biological consequences

While genetics has been successful in identifying susceptibility loci in complex diseases, the subsequent functional studies have been less successful in explaining how these genes affect disease. However, we expect this will be partly resolved in the near future as genetics will capitalize on other high-throughput approaches that are now being adopted in the field of biology. Through careful integration of these data-

Table 2: Overview of 16 diseases included in routine screening of newborns in the Netherlands

Disorder	Symptoms	Treatment
3-methylcrotonyl-CoA carboxylase deficiency	Feeding difficulties, vomiting, diarrhea, lethargy, hypotonia, development delay, seizures, coma	Low-protein diet, supplements
Biotinidase deficiency	Seizures, developmental delay, eczema, hearing loss	Biotin supplements
Congenital adrenal hyperplasia	Ambiguous genitalia, infertility, hypertension, vomiting, early pubic hair	Hormone and salt administration
Congenital hypothyroidism	Excessive sleeping, poor muscle tone, jaundice	Thyroxine administration
Galactosemia	Ataxia, speech deficits, dysmetria, premature ovarian failure, cataract	Lactose and galactose intake restriction
Glutaric aciduria type I	Brain damage, mental retardation, spasms, jerking, rigidity or decreased muscle tone and muscle weakness	Carnitine administration, restriction of protein (tryptophan, lysine) intake
HMG-CoA-lyase deficiency	Vomiting, dehydration, lethargy, convulsions, and coma	IV glucose and bicarbonate administration, restriction of protein (leucine) intake
Holocarboxylase synthase deficiency	Immunodeficiency diseases, feeding difficulties, breathing problems, skin rash, alopecia, lethargy	Biotin supplements
Homocystinuria	Thrombosis, cardiovascular failure	Vitamin B6
Isovaleric acidemia	Distinctive odor, sweety feets, feeding difficulties, vomiting, seizures, lethargy, coma	Protein intake restriction, glycine and carnitine administration
Long-chain hydroxyacyl-CoA dehydrogenase deficiency	lethargy, hypoglycemia, hypotonia, liver damage, heart damage, retina damage, muscle damage, peripheral neuropathy	Glucose administration
Maple syrup urine disease	Neurological damage, vomiting, dehydration, lethargy, hypotonia, seizures, ketoacidosis	Leucine, isoleucine and valine restriction, protein supplements
MCAD deficiency	hypoglycemia, hyperammonemia, vomiting, lethargy, neurological damage	Avoidance of fasting
Phenylketonuria	Seizures, microcephaly, impairment of cerebral function, mental retardation	Phenylalanine intake restriction
Sickle-cell disease	Ischemia, pain, tissue damage, acute chest syndrome	Zinc administration
Very-long-chain acyl-CoA dehydrogenase deficiency	Hypoketotic hypoglycemia, hepatocellular disease, cardiomyopathy, fatal encephalopathy	Avoidance of fasting, long-chain fatty acid intake restriction, carnitine supplements

Deletion

40,000,000 bp

sets (such as genomics and proteomics), considerable gains in the knowledge on functional consequences of genetic variation are likely to be made.

An important source of information is already building up: genetical genomics has been shown to be successful, as a considerable number of human genes have been identified whose expression is under strong genetic control[61; 62; 91-93]. Although only 2,500 unique individuals have been analyzed so far, many expression effects have already been observed.

We think that new (large-scale) genotyping projects should also consider systematically studying the gene expression for all genotyped samples. Although various problems remain to be resolved (such as tissue specificity of gene expression and RNA degradation issues), the extra cost of performing these analyses is modest compared to the cost of solely genotyping these samples. The fundamental knowledge likely to be provided will be of value to many researchers. These datasets will not only help to identify genes whose expression is under tight genetic control by variants in their vicinity, but analogous to findings in plants and mice, they will help to uncover new molecular pathways.

With the availability of more high-throughput techniques, such as proteomics and arrays for assessing genome-wide methylation, we expect the power to discover these pathways to increase even more. If experiments are properly performed using sufficient samples, many unknown pathways are likely to be unveiled, providing new avenues of investigation for molecular biologists.

Unintended consequences of our findings: commercial genetic testing

Dutch healthcare routinely uses screening and diagnostics for various genetic diseases. Postnatal screening is now systematically performed for16 genetic disorders, as determined by the Health Council of the Netherlands. These diseases have been included in routine screening under the principles outlined in 1968 by the World Health Organization[94]; all have severe consequences (table 2) if they remain untreated, but early intervention can either prevent their development or decrease their severity. Prenatal screening is also frequently performed. As the average age at which women become pregnant has increased steadily in the Netherlands over the past few decades, the chances of having a child with chromosomal abnormalities has increased as well. Dutch women are now systematically offered a combination test to assess for this (through echography, chorionic villus sampling, or amniocentesis). Apart from this, if there are other indications that a fetus might suffer from a severe genetic predisposition (such as a family history of Duchenne's muscular dystrophy), prenatal screening and diagnosis can be performed for a particular condition. Likewise, if people suspect they may have a familial predisposition for a certain disorder they can request genetic counseling. Based on the findings of a clinical geneticist, intervention or treatment is sometimes possible and advice on family planning can be given.

Clinical geneticists, who have been extensively trained to explain the ramifications of genetic findings to individuals are, by law, the only ones who may perform this counseling. It is important to realize that for many common disorders there is usually no black-and-white answer, but rather a probability or risk of developing the disease is determined. Assessing this risk can be very difficult (e.g. in breast cancer), as a trade-off has to be made between two potentially drastic courses (i.e. living with an increased risk of developing breast cancer or having breast amputation). As the eight clinical genetic centers in the Netherlands effectively had a monopoly on genetic testing and counseling, efficient quality control by the Dutch government was possible. The Health Council of the Netherlands ensured that patients were well informed and that well-articulated decisions were made.

The clinical genetics centers have, in the past, mainly provide services for rare Mendelian disorders that have severe consequences. However, a new era is opening up to the general public, in that commercial genetic tests that can also test for more complex disorders started to become available in 2006, allowing individuals to bypass the

traditional Dutch healthcare system. Initially these DNA tests only covered single gene disorders, but at the end of 2007 deCODEme and 23AndMe announced they could genotype hundreds of thousands of SNPs for $1,000 and use the results to provide risk estimates for dozens of complex diseases simultaneously (deCODEme is currently providing risk estimates for age-related macular degeneration, Alzheimer's disease, asthma, atrial fibrillation, breast cancer, celiac disease (utilizing our findings[1]), colorectal cancer, Crohn's disease, exfoliation glaucoma Xfg, multiple sclerosis, myocardial infarction, obesity, prostate cancer, psoriasis, restless legs, rheumatoid arthritis, type 1 diabetes and type 2 diabetes).

However, these predictions are difficult to interpret because, for each of these diseases, the genetic risk that is conferred by known loci is less than 10%. Given a realistic scenario of a disorder with a prevalence of 1% and five known susceptibility loci, each having a relative risk of 1.25, the individual's risk of developing this disease would theoretically range between 0.3% and 3.0%. How can one interpret these observations? A complication is that the odds ratios reported in the literature, used as the basis for estimating these risks, are currently likely to be overestimated, due to the 'winner's curse' phenomenon[95]. If a true-positive but small-effect disease locus is discovered, it is likely that chance favored its detection, so that in reality the actual odds ratio will be somewhat lower than that initially reported. Additionally, most of the currently associated SNPs tag a causal variant. These causal variants might exert a greater effect than the tag SNP, thereby complicating the calculation of accurate risk estimates even further. deCODEme's service agreement assumes these statistical aspects are clear to all of its clients: "*You acknowledge your understanding of genetic risk as a statistical measure that has implications derived from a large group of people with characteristics equivalent to yours but does not determine your chances for getting the corresponding disease*". However, given the small effects of the identified loci on the disease probabilities and the uncertainties of these probabilities for each of the diseases, it is extremely difficult, even for a clinical or statistical geneticist, to conclude anything.

The genetic test providers state that if you have an increased genetic risk, you might be able to compensate for this by reducing your environmental risk through exercising, losing weight or other changes in lifestyle. However, for some of the diseases that are offered for testing, the environmental factors are either unknown or can hardly be influenced (e.g. Alzheimer's disease or inflammatory bowel disease). Individuals might become concerned if they discover they have an increased risk of developing certain disorders and as a consequence they may visit their doctor more frequently, putting extra pressure on the healthcare system[96]. Genetic test results are likely to affect family relationships, because they may have consequences (to a lesser degree) for related individuals. Should individuals have the right to test themselves without even consulting their family members (let alone getting informed consent from them)?

Another relevant aspect is that under Dutch law, insurance companies have the right to ask whether individuals or their close relatives have been genetically tested for predispositions. This information can currently be used to accept or reject individuals for life insurance policies above € 160,000 or for total permanent disability insurance policies over € 32,000. It will be interesting to see how insurance companies react if a person replies that his/her brother has been genetically tested and it has been predicted he has a chance of 1.5% instead of the population average of 1% to develop a certain severe disorder.

Although these genetic tests can be used to assess ancestry (which is particularly interesting for many North Americans), the limited number of susceptibility loci and the difficulties of interpreting the calculated risks considerably restrict their usefulness. We envisage that in the coming years a higher proportion of the genetic variation for many diseases will be explained as more susceptibility loci are identified and new statistical frameworks employed (such as models that take epistasis into account). This will result in more accurate risk estimates, which will make the results from some of these tests

easier to interpret. As individuals can currently download all their genotyped SNPs (over 500,000), they will be able to assess these risks in the future by performing post-hoc analyses on their own data. Consequently, it will also enable the individuals to test for severe conditions that are potentially untreatable. If such genetic tests could predict accurately whether one would develop Alzheimer's disease or not, would one want to know? To prevent fear and anxiety, it can be argued that the tests offered should be confined to diseases for which intervention is possible. As such, we think the genetic testing companies should be (legally) restricted to provide individual genotype data.

We do think genetic tests that use much improved risk predictions in the future might actually contribute to better public health. For some strongly heritable diseases effective intervention is possible, enabling individuals to affect onset or expression of disease. To ensure its contribution to better public health we think the following guidelines should be imposed (ideally through jurisdiction) to prevent misinterpretation of genetic test results, unjustified fears, or legal abuse:

- To prevent overestimation of disease risk, the models that calculate risk should not be based on the odds ratios for risk loci reported by the papers that first identified them. Instead these odds ratios should solely be based on completely independent replications in unbiased cohorts.
- Risks should only be calculated for diseases for which the risk can be well interpreted and for which intervention is possible. Intervention should significantly affect onset or expression of disease, either through changes in lifestyle, medication or medical follow-up. Alternatively, family planning advice should enable people to make well-balanced decisions for their offspring.
- The outcomes of genetic testing should benefit public health and should minimize individuals' exposure to potential insurance problems.
- Individuals and all their family members need to be fully informed about the genetic test being taken. They need to be made aware of the uncertainties in risk prediction and interpretation, but also need to realize that science will progress quickly, enabling them to assess their risk for getting severe diseases in the future. They should also be informed of all the potential medical, familial and legal consequences.

Conclusions

The work we have carried out in the past four years has added to our understanding of complex diseases, specifically celiac disease and sporadic ALS. We have developed statistical methods that will help identify new disease pathways and susceptibility loci. However, more professional attention should be devoted to the societal implications of findings from genetic research, so that the progress made in genetics will indeed help to improve individual welfare and public health in general.

References

1 van Heel DA, Franke L, Hunt KA, Gwilliam R, Zhernakova A, Inouye M, Wapenaar MC, *et al.* (2007) A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nature genetics 39:827-829

2 van Es MA, van Vught PW, Blauw HM, Franke L, Saris CG, Van den Bosch L, de Jong SW, de Jong V, Baas F, van't Slot R, Lemmens R, Schelhaas HJ, Birve A, Sleegers K, Van Broeckhoven C, Schymick JC, Traynor BJ, Wokke JH, Wijmenga C, Robberecht W, Andersen PM, Veldink JH, Ophoff RA, van den Berg LH (2008) Genetic variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis. Nature genetics 40:29-31

3 Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. American journal of human genetics 78:1011-1025

4 Sutton WS (1903) The chromosomes in heredity. Biol Bull 4:231-251

5 Vries HD (1904) The Evidence of Evolution. Science (New York, NY 20:395-401

6 Boveri T (1904) Noch ein Wort Über Seeigelbastarde. Development Genes and Evolution 17:521-525

7 Morgan TH (1911) The Origin of Five Mutations in Eye Color in Drosophila and Their Modes of Inheritance. Science (New York, NY 33:534-537

8 Morgan TH (1911) The Origin of Nine Wing Mutations in Drosophila. Science (New York, NY 33:496-499

9 Bell J, Haldane JB (1937) The Linkage between the Genes for Colour-Blindness and Haemophilia in Man. Proceedings of the Royal Society of London Series B, Biological Sciences 123:119-150

10 Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171:737-738

11 Kan YW, Dozy AM (1978) Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation. Proceedings of the National Academy of Sciences of the United States of America 75:5631-5635

12 Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. American journal of human genetics 32:314-331

13 Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, Ottina K, Wallace MR, Sakaguchi AY, *et al.* (1983) A polymorphic DNA marker genetically linked to Huntington's disease. Nature 306:234-238

14 Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21:523-542

15 Ott J (1974) Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. American journal of human genetics 26:588-597

16 Lathrop GM, Lalouel JM (1985) Efficiency of recombination estimates using two-and three-point linkage data. Progress in clinical and biological research 194:97-102

17 Lathrop GM, Lalouel JM, Julier C, Ott J (1985) Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. American journal of human genetics 37:482-498

18 Jeffreys AJ, Wilson V, Thein SL (1985) Individual-specific 'fingerprints' of human DNA. Nature 316:76-79

19 Jeffreys AJ, Wilson V, Thein SL (1985) Hypervariable 'minisatellite' regions in human DNA. Nature 314:67-73

20 Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harbor symposia on quantitative biology 51 Pt 1:263-273

21 Litt M, Luty JA (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. American journal of human genetics 44:397-401

22 Weber JL, May PE (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. American journal of human genetics 44:388-396

23 Tabor HK, Risch NJ, Myers RM (2002) Candidate-gene approaches for studying complex genetic traits: practical considerations. Nature reviews 3:391-397

24 Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nature genetics 33 Suppl:228-237

25 Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC (1990) Linkage of early-onset familial breast cancer to chromosome 17q21. Science (New York, NY 250:1684-1689

26 Lander ES (1996) The new genomics: global views of biology. Science (New York, NY 274:536-539

27 Saunders AM, Strittmatter WJ, Schmechel D, George-Hyslop PH, Pericak-Vance MA, Joo SH, Rosi BL, Gusella JF, Crapper-MacLachlan DR, Alberts MJ, *et al.* (1993) Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. Neurology 43:1467-1472

28 Bertina RM, Koeleman BP, Koster T, Rosendaal FR, Dirven RJ, de Ronde H, van der Velden PA, Reitsma PH (1994) Mutation in blood coagulation factor V associated with resistance to activated protein C. Nature 369:64-67

29 Dean M, Carrington M, Winkler C, Huttley GA, Smith MW, Allikmets R, Goedert JJ, Buchbinder SP, Vittinghoff E, Gomperts E, Donfield S, Vlahov D, Kaslow R, Saah A, Rinaldo C, Detels R, O'Brien SJ (1996) Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study. Science (New York, NY 273:1856-1862

30 Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science (New York, NY 273:1516-1517

31 Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, *et al.* (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921

32 Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, *et al.* (2001) The sequence of the human genome. Science (New York, NY 291:1304-1351

33 International HapMap Consortium. (2005) A haplotype map of the human genome. Nature 437:1299-1320

34 Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851-861

35 Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Sato H, Hori M, Nakamura Y, Tanaka T (2002) Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nature genetics 32:650-654

36 Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement factor H polymorphism in age-related macular degeneration. Science (New York, NY 308:385-389

37 Pe'er I, de Bakker PI, Maller J, Yelensky R, Altshuler D, Daly MJ (2006) Evaluating and improving power in whole-genome association studies using fixed marker sets. Nature genetics 38:663-667

38 Barrett JC, Cardon LR (2006) Evaluating coverage of genome-wide association studies. Nature genetics 38:659-662

39 Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, Abraham C, Regueiro M, Griffiths A, Dassopoulos T, Bitton A, Yang H, Targan S, Datta LW, Kistner EO, Schumm LP, Lee AT, Gregersen PK, Barmada MM, Rotter JI, Nicolae DL, Cho JH (2006) A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science (New York, NY 314:1461-1463

40 Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, *et al.* (2007) Replicating genotype-phenotype associations. Nature 447:655-660

41 Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, *et al.* (2006) Global variation in copy number in the human genome. Nature 444:444-454

42 Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, Murthy KK, Rovin BH, Bradley W, Clark RA, Anderson SA, O'Connell R J, Agan BK, Ahuja SS, Bologna R, Sen L, Dolan MJ, Ahuja SK (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science (New York, NY 307:1434-1440

43 Aitman TJ, Dong R, Vyse TJ, Norsworthy PJ, Johnson MD, Smith J, Mangion J, Roberton-Lowe C, Marshall AJ, Petretto E, Hodges MD, Bhangal G, Patel SG, Sheehan-Rooney K, Duda M, Cook PR, Evans DJ, Domin J, Flint J, Boyle JJ, Pusey CD, Cook HT (2006) Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. Nature 439:851-855

44 Fellermann K, Stange DE, Schaeffeler E, Schmalzl H, Wehkamp J, Bevins CL, Reinisch W, Teml A, Schwab M, Lichter P, Radlwimmer B, Stange EF (2006) A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. American journal of human genetics 79:439-448

45 Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, *et al.* (2007) The diploid genome sequence of an individual human. PLoS biology 5:e254

46 Egholm M, Srinivasan M, Wheeler DA, A. M, He W, Chen Y, Makhijani V, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Chinault AC, Yuan Y, Jiang H, Song X, Liu Y, Nazareth L, Scherer S, Lupski JR, M. MD, Margulies M, Weinstock GM, Gibbs RA, M. RJ The Genome of James Dewey Watson. In preparation

47 Jansen RC, Nap JP (2001) Genetical genomics: the added value from segregation. Trends Genet 17:388-391

48 Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, Friend SH (2003) Genetics of gene expression surveyed in maize, mouse and man. Nature 422:297-302

49 Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. Science (New York, NY 296:752-755

50 Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L (2003) Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nature genetics 35:57-64

51 Storey JD, Akey JM, Kruglyak L (2005) Multiple locus linkage analysis of genomewide expression in yeast. PLoS biology 3:e267

52 Petretto E, Mangion J, Dickens NJ, Cook SA, Kumaran MK, Lu H, Fischer J, Maatz H, Kren V, Pravenec M, Hubner N, Aitman TJ (2006) Heritability and tissue specificity of expression quantitative trait loci. PLoS genetics 2:e172

53 Petretto E, Mangion J, Pravanec M, Hubner N, Aitman TJ (2006) Integrated gene expression profiling and linkage analysis in the rat. Mamm Genome 17:480-489

54 Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, Maciver F, Mueller M, Hummel O, Monti J, Zidek V, Musilova A, Kren V, Causton H, Game L, Born G, Schmidt S, Muller A, Cook SA, Kurtz TW, Whittaker J, Pravenec M, Aitman TJ (2005) Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nature genetics 37:243-253

55 Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P, Cookson WO, Taylor MS, Rawlins JN, Mott R, Flint J (2006) Genome-wide genetic association of complex traits in heterogeneous stock mice. Nature genetics 38:879-887

56 Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P, Edwards S, Phillips JW, Sachs A, Schadt EE (2004) Genetic inheritance of gene expression in human cell lines. American journal of human genetics 75:1094-1105

57 Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG (2004) Genetic analysis of genome-wide variation in human gene expression. Nature 430:743-747

58 Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, Morley M, Spielman RS (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. Nature genetics 33:422-425

59 Cheung VG, Spielman RS (2002) The genetics of variation in gene expression. Nature genetics 32 Suppl:522-525

60 Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT (2005) Mapping determinants of human gene expression by regional and genome-wide association. Nature 437:1365-1369

61 Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, Akey JM (2007) Gene-expression variation within and among human populations. American journal of human genetics 80:502-509

62 Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavare S, Deloukas P, Hurles ME, Dermitzakis ET (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science (New York, NY 315:848-853

63 Keurentjes JJ, Fu J, Terpstra IR, Garcia JM, van den Ackerveken G, Snoek LB, Peeters AJ, Vreugdenhil D, Koornneef M, Jansen RC (2007) Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. Proceedings of the National Academy of Sciences of the United States of America 104:1708-1713

64 Keurentjes JJ, Fu J, de Vos CH, Lommen A, Hall RD, Bino RJ, van der Plas LH, Jansen RC, Vreugdenhil D, Koornneef M (2006) The genetics of plant metabolism. Nature genetics 38:842-849

65 Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, Heath S, Depner M, von Berg A, Bufe A, Rietschel E, Heinzmann A, Simma B, Frischer T, Willis-Owen SA, Wong KC, Illig T, Vogelberg C, Weiland SK, von Mutius E, Abecasis GR, Farrall M, Gut IG, Lathrop GM, Cookson WO (2007) Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature 448:470-473

66 Hunt KA, Zhernakova A, Turner G, Heap G, Franke L, Bruinenberg M, Romanos J, *et al.* (2008) Novel coeliac disease genetic risk loci with links to adaptive immunity. Nature genetics In Press

67 Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661-678

68 Zhernakova A, Alizadeh BZ, Bevova M, van Leeuwen MA, Coenen MJ, Franke B, Franke L, Posthumus MD, van Heel DA, van der Steege G, Radstake TR, Barrera P, Roep BO, Koeleman BP, Wijmenga C (2007) Novel association in chromosome 4q27 region with rheumatoid arthritis and confirmation of type 1 diabetes point to a general risk locus for autoimmune diseases. American journal of human genetics 81:1284-1288

69 Cronin S, Berger S, Ding J, Schymick JC, Washecka N, Hernandez DG, Greenway MJ, Bradley DG, Traynor BJ, Hardiman O (2007) A genome-wide association study of sporadic ALS in a homogenous Irish population. Hum Mol Genet

70. Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, Saemundsen E, Stefansson H, Ferreira MA, Green T, Platt OS, Ruderfer DM, Walsh CA, Altshuler D, Chakravarti A, Tanzi RE, Stefansson K, Santangelo SL, Gusella JF, Sklar P, Wu BL, Daly MJ (2008) Association between microdeletion and microduplication at 16p11.2 and autism. The New England journal of medicine 358:667-675

71 Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, *et al.* (2007) Strong association of de novo copy number mutations with autism. Science (New York, NY 316:445-449

72 Elbers CC, Onland-Moret NC, Franke L, Niehoff AG, van der Schouw YT, Wijmenga C (2007) A strategy to search for common obesity and type 2 diabetes genes. Trends in endocrinology and metabolism: TEM 18:19-26

73 Torekov SS, Larsen LH, Andersen G, Albrechtsen A, Glumer C, Borch-Johnsen K, Jorgensen T, Hansen T, Pedersen O (2006) Variants in the 5' region of the neuropeptide Y receptor Y2 gene (NPY2R) are associated with obesity in 5,971 white subjects. Diabetologia 49:2653-2658

74 Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J (2007) QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic acids research 35:2013-2025

75 Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF,

Hakonarson H, Bucan M (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome research 17:1665-1674

76 Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nature genetics 39:906-913

77 Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature genetics 37:413-417

78 Service RF (2006) Gene sequencing. The race for the $1000 genome. Science (New York, NY 311:1544-1546

79 Cohen J, Pertsemlidis A, Kotowski IK, Graham R, Garcia CK, Hobbs HH (2005) Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. Nature genetics 37:161-165

80 Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science (New York, NY 305:869-872

81 Cohen JC, Pertsemlidis A, Fahmi S, Esmail S, Vega GL, Grundy SM, Hobbs HH (2006) Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proceedings of the National Academy of Sciences of the United States of America 103:1810-1815

82 Romeo S, Pennacchio LA, Fu Y, Boerwinkle E, Tybjaerg-Hansen A, Hobbs HH, Cohen JC (2007) Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. Nature genetics 39:513-516

83 Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, Weinstock GM, Gibbs RA (2007) Direct selection of human genomic loci by microarray hybridization. Nature methods 4:903-905

84 Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, McCombie WR (2007) Genome-wide in situ exon capture for selective resequencing. Nature genetics 39:1522-1527

85 Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME (2007) Microarray-based genomic selection for high-throughput resequencing. Nature methods 4:907-909

86 Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig CJ, Dahl F, Gao Y, Church GM, Shendure J (2007) Multiplex amplification of large sets of human exons. Nature methods 4:931-936

87 Parameswaran P, Jalili R, Tao L, Shokralla S, Gharizadeh B, Ronaghi M, Fire AZ (2007) A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. Nucleic acids research 35:e130

88 Hoffmann C, Minkah N, Leipzig J, Wang G, Arens MQ, Tebas P, Bushman FD (2007) DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. Nucleic acids research 35:e91

89 Binladen J, Gilbert MT, Bollback JP, Panitz F, Bendixen C, Nielsen R, Willerslev E (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. PLoS ONE 2:e197

90 Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nature methods Early Access

91 Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M, Lathrop GM, Abecasis GR, Cookson WO (2007) A genome-wide association study of global gene expression. Nature genetics 39:1202-1207

92 Göring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole SA, Jowett JB, Abraham LJ, Rainwater DL, Comuzzie AG, Mahaney MC, Almasy L, MacCluer JW, Kissebah AH, Collier GR, Moses EK, Blangero J (2007) Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nature genetics 39:1208-1216

93 Myers AJ, Gibbs JR, Webster JA, Rohrer K, Zhao A, Marlowe L, Kaleem M, Leung D, Bryden L, Nath P, Zismann VL, Joshipura K, Huentelman MJ, Hu-Lince D, Coon KD, Craig DW, Pearson JV, Holmans P, Heward CB, Reiman EM, Stephan D, Hardy J (2007) A survey of genetic human cortical gene expression. Nature genetics 39:1494-1499

94 Wilson JM, Jungner YG (1968) Principles and practice of mass screening for disease. Boletin de la Oficina Sanitaria Panamericana 65:281-393

95 Ioannidis JP (2003) Genetic associations: false or true? Trends in molecular medicine 9:135-138

96 Hunter DJ, Khoury MJ, Drazen JM (2008) Letting the genome out of the bottle--will we get our wish? The New England journal of medicine 358:105-107

Samenvatting

Een groot aantal ziekten zijn erfelijk. Hierbij valt te denken aan vrij zeldzame ziekten als cystic fibrose (taaislijmziekte) of amyotrofe laterale sclerose, maar ook aan veelvoorkomende ziekten zoals borstkanker, atherosclerose (aderverkalking), reuma, diabetes (suikerziekte), astma en coeliakie (glutenintolerantie).

Genetisch onderzoek richt zich onder andere op het opsporen van de erfelijke factoren die bijdragen aan het ontstaan van deze ziekten. In 1983 werd het eerste grote succes geboekt: Er werd een locus op het DNA gevonden waar een mutatie moest liggen die de ziekte van Huntington veroorzaakt. Daarna volgden successen elkaar snel op en werden er voor een groot aantal, voornamelijk zeldzame aandoeningen, ziekteverwekkende loci gevonden in het DNA.

Pas in 1990 werd voor het eerst een afwijking in een veelvoorkomende ziekte gevonden, te weten borstkanker. De mutatie kan een deel van de gevallen van borstkanker verklaren. Helaas bleek het opsporen van de precieze genetische risicofactoren voor andere ziekten die veel voorkomen toch aanzienlijk moeilijker.

Belangrijke verklaringen voor het uitblijven van succes voor deze ziekten waren de technische mogelijkheden van die tijd en de manier waarop genetische studies werden opgezet: Aangenomen werd dat telkens één of enkele mutaties een ziekte verklaart. Uitgaande van deze vooronderstelling, werd per chromosoom systematisch op een aantal loci gekeken of er verschillen zichtbaar waren tussen patiënten en controles. Het beperkte succes van dit 'koppelingsonderzoek' leidde tot het besef dat er vermoedelijk een groot aantal verschillende genetische risicofactoren een rol spelen. Daarom was een andere insteek wenselijk.

Echter, methoden om dit daadwerkelijk te doen waren nog niet voorhanden, doordat nog onvoldoende inzicht in het menselijk genoom aanwezig was. Dit veranderde in 2001, toen het menselijke genoom systematisch in kaart was gebracht. Deze mijlpaal maakte het mogelijk om systematisch te gaan kijken naar de normale genetische verschillen tussen gezonde mensen. Al snel werd het duidelijk dat er op miljoenen plekken in het DNA kleine verschillen (SNPs) tussen mensen aanwezig zijn. Dit inzicht lag ten grondslag aan de ontwikkeling van zo-

Begrippenlijst

Genoom	Erfelijke informatie, beschrijft het grootste deel van de onderdelen van een cel
DNA	Drager van de erfelijk informatie
RNA	Stap tussen DNA en eiwit
Chromosoom	DNA is normaliter verdeeld over 46 verschillende chromosomen per menselijke cel
Locus	Plaats op het DNA
Mutatie	Verandering in het DNA
SNP	Single nucleotide polymorphism (spreek uit 'snip'). Verandering in het DNA van één basepaar (nucleotide)

genaamde ‘DNA chips’ in 2006. Hiermee kan in één keer naar honderdduizenden SNPs per individu worden gekeken *(dit proefschrift geeft op iedere pagina een deel van mijn eigen DNA weer)*. Aangezien dit ongeveer duizend keer zoveel informatie per individu opleverde dan koppelingsonderzoek, werd de kans op het succesvol vinden van nieuwe ziekteverwekkende loci hoger.

Dit proefschrift beschrijft allereerst nieuwe statistische methoden om genetisch onderzoek uit te voeren en deze DNA chips te analyseren. Tevens zijn een tweetal studies beschreven waarbij deze DNA chips zijn ingezet. Tenslotte wordt stilgestaan bij toekomstige ontwikkelingen en maatschappelijke consequenties van deze resultaten.

In hoofdstuk 2 wordt de ontwikkeling van een netwerk van relaties tussen genen en eiwitten beschreven. Het menselijke DNA codeert voor naar schatting 20.000 verschillende genen. Deze genen worden meestal vertaald in eiwitten die voorzien in verschillende functies in een cel. Sommige eiwitten werken samen in dezelfde biologische processen. Verstoringen in deze biologische processen kunnen leiden tot ziekte. Wij veronderstelden dat mutaties in verschillende genen binnen zo’n biologisch proces daarom tot eenzelfde ziekte kunnen leiden. Op basis van deze hypothese is het netwerk toegepast op ziekte loci die gevonden zijn met behulp van koppelingsonderzoek. Deze loci bevatten vaak een groot aantal verschillende genen, waarbij normaliter verondersteld wordt dat één van deze genen een mutatie zal bevatten. Aan de hand van simulaties hebben we laten zien dat onze methode met behulp van het gen netwerk in staat is vaker dan verwacht het gemuteerde ziekte gen correct aan te wijzen. Deze observatie bevestigt onze hypothese dat de verschillende genetische risicofactoren binnen één enkele ziekte vaak in genen zitten die vergelijkbare biologische functies hebben.

Hoofdstuk 3 beschrijft een nieuwe statistische methode waarmee een groot aantal verschillende maar vaak voorkomende deleties (structurele varianten) zijn geïdentificeerd in de Nederlands en Engelse populatie. Het toegepaste algoritme stelde ons in staat ongeveer 700 deleties te vinden die over het algemeen kleiner zijn dan eerder gevonden structurele varianten. Bekend is

dat sommige structurele varianten ook risicofactoren voor ziekten zijn. Een analyse in Engelse coeliakie patiënten en gezonde controles identificeerde *CSF1R* als een gen waarin een deletie mogelijk is geassocieerd met coeliakie. Vervolgonderzoek zal moeten uitwijzen wat de rol van dit gen in coeliakie is. Daarnaast verwachten we dat deze methode bij gebruikmaking van nieuwe DNA chips meer structurele varianten zal vinden. De momenteel uitgevoerde analysen maakten gebruik van DNA chips, waarbij het ontwerp onbedoeld het vinden van deze structurele varianten bemoeilijkt. De nieuwe DNA chips kennen dit probleem minder en bemonsteren daarnaast meer plekken op het DNA, waardoor een groot aantal, niet eerder beschreven structurele varianten, gevonden zullen worden.

Hoofdstuk 4 beschrijft een nieuwe manier om ziekten te bestuderen waarvan de genetische basis nu nog grotendeels onduidelijk is, maar waarbij er aanwijzingen zijn dat er bij patiënten vaak delen van hun DNA missen (deleties) of vaker dan gewoonlijk voorkomen (gedupliceerd). Wij veronderstellen dat deze mutaties effect hebben op het functioneren van de genen die in deze loci liggen. Aangezien in deze loci vaak verscheidene genen liggen zullen dus verschillende biologische processen aangedaan zijn. In autisme hebben we deze hypothese getoetst op basis van dertien eerder beschreven loci. Voor ieder locus hebben we per gen bepaald welke klinische symptomen door mutaties veroorzaakt kunnen worden. Vervolgens hebben we gekeken of bepaalde symptomen door mutaties in meer loci dan verwacht veroorzaakt kunnen worden. Aangezien symptomen die bekend zijn samen te gaan met autisme, zoals gezichtsafwijkingen en epilepsie, kwamen hieruit naar voren, onderschrijft dit bovenstaande hypothese.

Hoofdstuk 5 beschrijft een analyse van de effecten die genetische variatie heeft op de RNA expressie van genen. Met behulp van DNA en RNA chips hebben we een groot aantal genen geïdentificeerd die beïnvloed worden door veelvoorkomende genetische varianten. Deze soort studies bevinden zich nog in een vroeg stadium waardoor het pas zeer recentelijk is gebleken dat de invloed van genetische variatie op gen expressie soms verkeerd is geïnterpreteerd. Allereerst blijkt voor een aanzienlijk deel van de genen dat het hier vals positieve bevindingen be-

treft. Voor verscheidene andere genen blijkt dat de regulatie complexer is dan verondersteld. Daarnaast blijkt dat voor sommige genetische varianten geldt dat deze leiden tot verschuivingen in verhoudingen van geproduceerde gen-varianten (splice varianten). De komst van nieuwe *high-throughput sequencing* technologieën zal het in de nabije toekomst mogelijk maken beter naar deze regulatie te kijken.

Hoofdstuk 6 beschrijft een studie met behulp van DNA chips in 778 Engelse coeliakie patiënten en 1.422 gezonde controles. Door middel van een nieuwe statistische methode voor het toekennen van genotypes, leidde deze studie tot de identificatie van een nieuw ziekteverwekkend locus waarin onder meer de genen *IL2* en *IL21* liggen. Deze interleukines spelen een belangrijke rol bij de ontstekingsreactie die kenmerkend is voor coeliakie: de Th1 adaptieve immuunreactie. Recentelijk is deze studie vervolgd in een grotere groep patiënten en controles. Dit heeft geleid tot de identificatie van nog eens zeven loci, waarvan zes loci genen bevatten met een bekende immunologische functie. Middels meta-analyses en de inzet van nieuwe DNA chips is het de verwachting dat extra loci een rol zullen blijken te spelen in coeliakie. Daarnaast zal met het behulp van de eerder genoemde nieuwe sequencing technologieën vermoedelijk mogelijk worden de causale variant op te sporen.

Hoofdstuk 7 beschrijft een studie met behulp van dezelfde DNA chips in 461 Nederlandse amyotrofe laterale sclerose patiënten en 450 gezonde controles. In deze studie werd een nieuw ziekteverwekkend locus gevonden waarin het gen *DPP6* ligt. *DPP6* komt voornamelijk tot expressie in het brein en beïnvloedt de biologische activiteit van neuropeptides. Verschillen in *DPP6* expressie zijn daarnaast in verband gebracht met schade aan het ruggenmerg in ratten. Deze resultaten bieden concrete handvatten voor vervolgonderzoek en bieden in de toekomst misschien aanknopingspunten voor therapeutische interventie.

In hoofdstuk 8 wordt een evaluatie gegeven van onze behaalde resultaten. Daarnaast worden verwachtingen voor toekomstig genetisch onderzoek geschetst. Afgesloten wordt met een discussie omtrent de inzet van DNA chips bij het voorspellen van individuele risico's voor het ontwikkelen van ziekte.

De belangrijkste conclusies van dit proefschrift zijn:

- De inzet van moleculaire gen netwerken in genetisch onderzoek kan helpen bij de identificatie van nieuwe genetische risicofactoren *(hoofdstuk 2)*
- Structurele genetische variatie is wijdverspreid in het humane genoom. De inzet van DNA oligonucleotide chips die niet gebiased zijn tegen structurele varianten zal leiden tot een beter en completer inzicht van structurele variatie *(hoofdstuk 3)*
- Syndromale ziekten met een klinisch variabel spectrum, zoals autisme, kunnen profijt hebben van een systematische genotype-fenotype analyse *(hoofdstuk 4)*
- Genetische variatie heeft een effect op gen expressie. Deze regulatie kan echter complex van karakter zijn *(hoofdstuk 5)*
- Het *IL2/IL21* locus is geassocieerd met coeliakie *(hoofdstuk 6)*
- Het *DPP6* locus is geassocieerd met amyotrofe laterale sclerose *(hoofdstuk 7)*
- De recente beschikbaarheid van DNA oligonucleotide chips die voor individuen en verzekeraars betaalbaar zijn geeft aanleiding tot een maatschappelijke discussie over de wenselijkheid van de inzet van deze chips in een vroegtijdig stadium *(hoofdstuk 8)*.

Summary

Many diseases are heritable. Among these are quite rare disorders such as cystic fibrosis or amyotrophic lateral sclerosis, but also common diseases such as breast cancer, atherosclerosis, rheumatoid arthritis, diabetes, asthma and celiac disease.

One of the aims of genetic research is to identify the genetic factors that contribute to these diseases. The first success in human disease was in 1983: a locus in the DNA was identified that had to contain a mutation that caused Huntington's disease. After this first success, many new disease-causing loci were identified in predominantly rare disorders.

In 1990 a milestone was achieved for common diseases: a mutation was found that was responsible for breast cancer in a minority of cases. However, the identification of the genetic risk factors for many other common diseases turned out to be more difficult.

Important explanations for the lack of success for these diseases were the technical possibilities available at the time and the way the genetic studies had been designed. It had been assumed that there were only a few causative mutations for each of these diseases. Based on this assumption, various loci in each chromosome were assessed systematically for evidence of whether there was a mutation in or near any of these loci. The limited success of this 'linkage analysis' approach implied that, for many diseases, it was likely that numerous different genetic risk factors could be playing a role. Alternative strategies were proposed to detect these.

However, these strategies could not yet be employed, because there was only limited insight into the human genome. This changed in 2001, when the Human Genome Project was completed and the full genome mapped. This milestone permitted a systematic analysis of normal genetic variation among healthy individuals. It soon became apparent that there were millions of places in the DNA where there were SNPs, small differences between humans.

Glossary

Genome	Heriditary information, describes the majority of the parts that make up a cell
DNA	Encodes the heritable information
RNA	Step in between DNA and protein
Chromosome	DNA is usually distributed over 46 different chromosomes per human cell
Locus	Place on the DNA
Mutation	Alteration within the DNA
SNP	Single nucleotide polymorphism (pronounce as 'snip'). Change in the DNA of one single base pair (nucleotide)

This knowledge was the basis for 'DNA chips', becoming available in 2006. These chips can simultaneously assess hundreds of thousands of SNPs per individual. This results in approximately a thousand times more information than from linkage analysis, increasing the probability of successfully identifying risk factors.

This thesis describes new statistical methods developed to analyze these DNA chips and to perform genetic research. Additionally it describes two studies that have employed these DNA chips. Finally, attention is devoted to future developments and the societal consequences of these results.

Chapter 2 describes the development of a network of relationships between genes and proteins. Human DNA codes for approximately 20,000 different genes. These genes are usually translated into proteins that carry out most of the functions within a cell and some proteins act together in the same biological processes. Alterations in these biological processes can lead to disease. We thus assumed that mutations in various genes that play a role in the same biological processes could result in the same disorder. Based on this hypothesis we applied our network to susceptibility loci that had been identified through linkage analysis. These loci are often fairly large and contain multiple genes. It is assumed that usually only one of these genes will contain a mutation. Based on simulations we showed that our method correctly pinpointed the real disease gene more often than would be expected by chance. This observation corroborates our hypothesis that, for many diseases, the different genetic risk factors map within genes that have comparable biological functions.

Chapter 3 describes a new statistical method that identified many different but common deletions (structural variants) through an analysis of Dutch and English samples. Our algorithm enabled us to identify approximately 700 deletions that on average are smaller than those found previously. It is known that these structural variants can also be risk factors for diseases. An analysis in English celiac disease patients and

healthy controls identified *CSF1R* as a gene that mapped within a deletion associated with celiac disease. Follow-up research is needed to determine its role in celiac disease. We expect our method to identify many more structural variants by employing new DNA chips. The current analyses were performed on DNA chips that were actually biased against finding these structural variants. As newer DNA chips are less biased and also query many more loci, we expect many more, new structural variants to be identified in the near future.

Chapter 4 describes a novel method to study diseases for which the genetic basis is still mostly unclear, but for which there is evidence that patients often miss parts of their DNA (deletions) or have multiple copies of parts of their DNA (duplications). We hypothesized that these mutations are likely to affect the function of the genes within these deletions or duplications. Since multiple genes usually map within these loci, it can be assumed that multiple biological processes are affected. We tested this hypothesis on thirteen previously described loci for autism. We determined which clinical symptoms could be caused by mutations within these genes for each loci per gene. We then assessed whether certain symptoms could be caused by mutations in more loci than expected. Various symptoms known to co-occur with autism were identified, such as craniofacial abnormalities and epilepsy. As such, these findings substantiate our hypothesis.

Chapter 5 describes an analysis of the effects of genetic variation on RNA gene expression. By using both DNA and RNA chips we identified many genes whose expression is affected by common genetic variants. These studies are still in their infancy, and it has only recently been shown that some of the observed effects have not been explained correctly. For a considerable proportion of these genes, the observed effect on expression was not real but rather due to technical artifacts. For other genes whose expression is truly affected by genetic variation, their regulation was found to be more

complex than originally anticipated. Some genetic variants can also lead to shifts in the ratios of gene variants (splice variants) produced in different ways. We expect more accurate and more complete insight will be achieved with the arrival of new high-throughput sequencing technologies.

Chapter 6 describes a DNA chip study in 778 English celiac disease patients and 1,422 healthy controls. By developing a new statistical method to assign genotypes to these samples, we were able to identify a new susceptibility locus that contains *IL2* and *IL21*. These interleukins play an important role in the immune response which is characteristic for untreated celiac disease patients (the Th1 adaptive immune response). Recently this study has been followed up in a larger cohort of patients and controls, which has led to the identification of seven additional loci. Six of these loci contain candidate genes with a known immunological function. Through meta-analyses and by employing new DNA chips, we should be able to identify various additional loci that play a role in celiac disease. New sequencing technologies will probably enable us to identify the real causal variants.

Chapter 7 describes a DNA chip study in 461 Dutch amyotrophic lateral sclerosis patients and 450 healthy controls. This study led to the identification of a new susceptibility locus that contains *DPP6*. This gene is predominantly expressed in brain and affects the biological activity of neuropeptides. Differences in *DPP6* expression have been associated with damage to the spinal cord in rats. These results provide new avenues for follow-up research and may provide clues for developing therapeutic interventions.

In Chapter 8 we evaluate our results and also provide some perspectives for future genetic research. Finally, we discuss the current usefulness of DNA chips in predicting individual risks for developing disease and the ethical issues surrounding commercial genetic screening services.

The main conclusions of this thesis are:

- The use of molecular gene networks in genetic research can help to identify new genetic risk factors *(chapter 2)*
- Structural genetic variants are widespread throughout the human genome. The use of DNA oligonucleotide chips that are not biased against these variants will lead to a better and more complete insight into structural variation *(chapter 3)*
- Syndromes with a broad clinical spectrum, such as autism, can benefit from a systematic genotype-phenotype analysis *(chapter 4)*
- Genetic variants can affect gene expression; however, its regulation can be complex *(chapter 5)*
- The *IL/IL21* locus is associated with celiac disease *(chapter 6)*
- The *DPP6* locus is associated with amyotrophic lateral sclerosis *(chapter 7)*
- The recent availability of affordable DNA oligonucleotide chips for individuals and insurance companies means society must debate the ethical and practical issues surrounding their use.

Dankwoord

Weliswaar staat op de omslag van dit proefschrift slechts de naam van één auteur, de bijdragen van anderen zijn minstens zo groot geweest. Het bewijs hiervoor vindt u deels in de infographics waarin de door mij ontvangen e-mail is weergegeven: Hieruit blijkt de grote inbreng en onschatbare waarde van velen. Ieder heeft vanuit een eigen invalshoek bijdragen geleverd aan de totstandkoming van dit proefschrift. Graag maak ik hierbij gebruik van de mogelijkheid een aantal mensen extra te bedanken.

Beste Cisca, dankzij jou heb ik de mogelijkheid gehad me te bekwamen in een specialisme wat ik ontzettend uitdagend, leuk en inspirerend vind. Ik heb verschrikkelijk veel van je geleerd en hoop dit de komende jaren te blijven doen in het rode Groningen!

Dear David, thank you for letting me work on the celiac genome-wide association study and the joy of Princes' street. Karen and Graham, I am looking forward to seeing you again in London.

Beste Leonard, bedankt dat je de mogelijkheid hebt geboden betrokken te zijn bij het ALS onderzoek. Beste Roel, dank voor de vele conversaties als ik weer eens voorbijliep in de gang op weg naar nog een koffie.

Beste Ritsert, met plezier kom ik iedere dinsdag naar Haren. Hopelijk mag ik nog een lange tijd langskomen bij jou en alle andere GBIC'ers voor intrigerende discussies.

Dear Lon, thank you for inviting me to come over to Oxford. I learned a lot from you and all other friends at the WTCHG.

Ik dank de leescommissie voor het kritisch beoordelen van dit proefschrift.

Ik ben het Celiac Disease Consortium en het Nationaal Regieorgaan Genomics dankbaar voor het financieren van mijn onderzoek.

Beste Michael van Es, Hylke, Christiaan, Paul en Jan: We hebben gezamenlijk veel geleerd over de nieuwe Illumina arrays en ALS, dank!

Beste Like: Ik vond het geweldig om aan het begin van mijn promotie met jou te mogen samenwerken. Beste Michael Egmont-Petersen: Hoofdstuk twee was zonder jouw Bayesiaanse statistische kennis en creativiteit absoluut niet mogelijk geweest.

Wouter van Gool, Tebbo, Iris, Kristel en Dirk-Jan: bedankt voor al jullie inspanningen als stagiaire. Ik heb veel van jullie geleerd en hoop dat het wederzijds is.

Beste Edwin, Arno en Rainer: Dank voor jullie advies op bioinformatica en text-mining gebied.

Beste Dalila, Bart, Dineke, Clara, Jonathan, Steven, Martine, Alienke, Marianna, Erica, Eric, Ruben, Bobby, Begoña, Roel, Jelena, Esther, Karen, Behrooz, Gosia, Martin en Sasha: Ik heb jullie leren kennen als een zeer kleurrijk gezelschap. Hopelijk heb ik jullie niet tot wanhoop gedreven als ik weer eens zat te drammen.

Beste Carolien en Yurii, dank voor de vele inspirerende gesprekken. Jullie ideeën waren instrumentaal voor hoofdstuk vier.

Beste Jacobine, Bert, Wouter Staal, Jacob, Ron, Emma en Herman: Hopelijk heeft onze samenwerking van onder andere hoofdstuk vijf niet geleid tot een DSM-IV etiket voor mij.

Beste Harm, dank voor je begeleiding tijdens mijn stage en onze immer prikkelende conversaties. Hierbij horen natuurlijk ook Yumas, Wim, Philip en Patrick: ik denk met genoegen terug aan de vele discussies die we met elkaar hebben gevoerd, waarbij we het opvallend vaak met elkaar oneens bleken te zijn. Excuses voor alle digitale overlast die ik jullie bezorgd heb.

Beste Harry: ik hoop, net als jij, na mijn vijfstigste nog vlot van een zwarte piste af te komen. Wat is je geheime recept? Zijn het de chocolade flensjes?

Allerbeste Jackie: velen hebben hun artikelen aan jou te danken. Voor steenkolen Engels heb jij me weten te behoeden, alsmede ook voor spellings- en grammatica- en typfouten en wanstaltige zinsconstructies, zoals deze.

Beste Alfons, met genoegen kijk ik terug op onze 'professionele' relatie: de vele concerten die we hebben bezocht, waarvan Oi Va Voi en Spinvis de toppers zijn geweest!

Beste Flip en Albertien: Sectie Senseo zeg ik met pijn in het hart vaarwel. Bedankt voor de gezellige tijd samen!

Gert en Thomas: Dankzij jullie inspanningen is dit boek geworden wat het is.

Aan al mijn collega's en Gert: Sorry voor het altijd te laat komen, te laat inleveren of te laat reageren!

Vrienden van de W55, AK19 en aanverwanten: In voor- en tegenspoed, altijd balkon of trap met een biertje. Bedankt voor jullie vriendschap!

Koen, Fleur, Sabine, Sander, Moniek: De duurgekochte nootjes heb ik graag met jullie gedeeld en het is fantastisch dat we nog altijd bij elkaar komen.

Lieve studievrienden: Jullie zijn net zo verbaasd als ik dat er in mij een bioloog school. Bedankt voor jullie hulp tijdens alle practica.

Roald, Erik en Wouter: Jullie stonden garant voor het nodige absurdisme de afgelopen jaren, *'of niet?'*. Het is een genoegen jullie te kennen.

Koen en Sytse: Geweldig dat jullie paranimf zijn. Koen, bedankt dat je bereid bent de verdediging inhoudelijk over te nemen, mocht ik onwel raken!

Lieve Gert en Lieke: Jullie zijn een top broer en zus! Lieve papa en mama: Dank jullie wel voor alle jaren zorg, liefde en kritiek op zijn tijd.

Lieve Lien: dank je wel voor onze prachtige Suzuki avonturen, de gedichten die ik waarschijnlijk toch nooit zal begrijpen, je steun, je optimisme en je humor. Jij bent echt de allerliefste!

List of publications

§ Equal contribution

22 **Franke L** [§] , Buizer-Voskamp JE [§], Staal WG, van Daalen2 E, Kemner C, Ophoff RA, Vorstman JAS, van Engeland H, Wijmenga C.
Systematic genotype-phenotype analysis of autism susceptibility loci implicates additional symptoms to co-occur with autism.
Submitted

21 **Franke L**, de Kovel CGH, Aulchenko YS, Trynka G, Zhernakova A, Hunt KA, Blauw HM, van den Berg LH, Ophoff RA, Deloukas P, van Heel DA, Wijmenga C.
Detection, Imputation and Association Analysis of Small Deletions and Null-alleles on Oligonucleotide Arrays.
American Journal of Human Genetics, in press

20 Zhernakova A, Festen EM, **Franke L**, Trynka G, Monsuur AJ, Bevova M, Nijmeijer RM, Heijmans R, van Heel DA, van Bodegraven AA, Stokkers PCF, Wijmenga C, Crusius JBA, Weersma RK.
Genetic analysis of the innate immune system identifies CARD9 and the IL18 receptor locus as susceptibility genes for both Crohn's disease and ulcerative colitis
American Journal of Human Genetics, in press

19 Hunt KA, **Franke L**, Deloukas P, Wijmenga C, van Heel DA.
No evidence in a large UK collection for celiac disease risk variants reported by a Spanish study.
Gastroenterology, in press

18 Hunt KA, Zhernakova A, Turner G, Heap G, **Franke L**, Bruinenberg M, Romanos J, Dinesen LC, Ryan AW, Panesar1 D, Gwilliam R, Takeuchi F, McLaren WM, Holmes GKT, Howdle PD, Walters JRF, Sanders DS, Playford RJ, Trynka G, Mulder CJJ, Mearin ML, Verbeek WHM, Trimble V, Stevens FM, O'Morain C, Kennedy NP, Kelleher D, Pennington DJ, Strachan DP, McArdle WL, Mein CA, Wapenaar MC, Deloukas P, McGinnis R, McManus R, Wijmenga C, van Heel DA.
Novel coeliac disease genetic risk loci with links to adaptive immunity.
Nature Genetics. 2008 Apr;40(4):395-402

17 Blauw HM, Veldink JH, van Es MA, van Vught PW, Saris CGJ, van der Zwaag B, **Franke L**, Burbach JPG, Wokke JH, Ophoff RA, van den Berg LH.
Genome-wide copy-number variation in amyotrophic lateral sclerosis
Lancet Neurology. 2008 Apr;7(4):319-26

16 van Vliet-Ostaptchouk JV, Onland-Moret NC, van Haeften TW, **Franke L**, Elbers CC, Shiri-Sverdlov N, van der Schouw YT, Hofker MH, Wijmenga C.
HHEX gene polymorphisms are associated with type 2 diabetes in the Dutch Breda cohort
European Journal of Human Genetics. 2008 Jan 30

15 **Franke L**[§], van Es MA[§], van Vught PW[§], Blauw HM[§], van den Bosch L, de Jong SW, de Jong V, Baas F, van 't Slot R, Lemmens R, Schelhaas HJ, Birve A, Sleegers K, van Broeckhoven C, Schymick JC, Traynor BJ, Wokke JH, Wijmenga C, Robberecht W, Andersen PM, Veldink JH, Ophoff RA, van den Berg LH. Genetic variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis
Nature Genetics. 2008 Jan;40(1):29-31

14 Zhernakova A, Alizadeh BZ, Bevova M, van Leeuwen MA, Coenen MJH, Franke B, **Franke L**, Posthumus MD, van Heel DA, van der Steege G, Radstake TRDJ, Barrera P, Roep BO, Koeleman BPC, Wijmenga C.
Novel association in chromosome 4q27 region to rheumatoid arthritis and confirmation to type 1 diabetes points to a general risk locus for autoimmune diseases
American J of Human Genetics. 2007 Dec;81(6):1284-8

13 **Franke L**[§], van Es MA[§], Van Vught PW[§], Blauw HM[§], Saris CG, Andersen PM, Van Den Bosch L, de Jong SW, van 't Slot R, Birve A, Lemmens R, de Jong V, Baas F, Schelhaas HJ, Sleegers K, Van Broeckhoven C, Wokke JH, Wijmenga C, Robberecht W, Veldink JH, Ophoff RA, van den Berg LH.
ITPR2 as a susceptibility gene in sporadic amyotrophic lateral sclerosis: a genome-wide association study.
Lancet Neurology. 2007 Oct;6(10):869-77.

40,000,000 bp

50,000,000 bp

Curriculum vitae

Lude Franke werd op 14 januari 1980 geboren te Oldenzaal. Na het eindexamen atheneum op de R.S.G. Noord-Oost Veluwe in 1998 begon hij aan zijn studie medische biologie aan de Universiteit Utrecht. In 2001 begon hij zijn grafisch ontwerpbureau Ludesign. Zijn stage vond plaats bij de afdeling Medische Genetica van het Universitair Medisch Centrum Utrecht, waar hij begeleid werd door Harm van Bakel en gen expressie en koppelingsonderzoek data integreerde. In 2002 studeerde hij af en vervolgde met een jaar filosofie aan de Radboud Universiteit, Nijmegen. Eind 2003 begon hij als AIO bij de afdeling Medische Genetica, Universitair Medisch Centrum Utrecht, waar hij zijn onderzoek vervolgde, wat uiteindelijk heeft geleid tot dit proefschrift. Tijdens zijn AIO periode was hij drie maanden te gast bij dr. Lon Cardon, Wellcome Trust Centre for Human Genetics, Oxford. Per 2008 is de auteur post-doc bij de afdeling Genetica van het Universitair Medisch Centrum Groningen en bij de afdeling Gastroenterologie van Queen Mary, University of London.

12 de Jong E, **Franke L**, Siebes A.
On the measurement of genetic interactions
AIP Conf. Proc. 2007 Sep 17; 940:16-25

11 van Heel DA, **Franke L**[§], Hunt KA[§], Gwilliam R[§], Zhernakova A, Inouye M, Wapenaar MC, Barnardo MC, Bethel G, Holmes GK, Feighery C, Jewell D, Kelleher D, Kumar P, Travis S, Walters JR, Sanders DS, Howdle P, Swift J, Playford RJ, McLaren WM, Mearin ML, Mulder CJ, McManus R, McGinnis R, Cardon LR, Deloukas P, Wijmenga C.
A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21.
Nature Genetics. 2007 Jul;39(7):827-9

10 Diosdado B, van Bakel H, Strengman E, **Franke L**, van Oort E, Mulder CJ, Wijmenga C, Wapenaar MC.
Neutrophil recruitment and barrier impairment in celiac disease: a genomic study. Clin Gastroenterol
Hepatology. 2007 May;5(5):574-581.e5

9 Elbers CC, Charlotte Onland-Moret N, **Franke L**, Niehoff AG, van der Schouw YT, Wijmenga C.
A strategy to search for common obesity and type 2 diabetes genes.
Trends Endocrinology Metab. 2007 Jan-Feb;18(1):19-26

8 Malik R, **Franke L**, Siebes A.
Combination of text-mining algorithms increases the performance.
Bioinformatics. 2006 Sep 1;22(17):2151-7

7 **Franke L**, Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C.
Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes.
American Journal of Human Genetics. 2006 Jun;78(6):1011-25

6 Vorstman JA, Staal WG, van Daalen E, van Engeland H, Hochstenbach PF, **Franke L**.
Identification of novel autism candidate regions through analysis of reported cytogenetic abnormalities associated with autism.
Molecular Psychiatry. 2006 Jan;11(1):1, 18-28

5 Monsuur AJ, de Bakker PI, Alizadeh BZ, Zhernakova A, Bevova MR, Strengman E, **Franke L**, van't Slot R, van Belzen MJ, Lavrijsen IC, Diosdado B, Daly MJ, Mulder CJ, Mearin ML, Meijer JW, Meijer GA, van Oort E, Wapenaar MC, Koeleman BP, Wijmenga C.
Myosin IXB variant increases the risk of celiac disease and points toward a primary intestinal barrier defect.
Nature Genetics. 2005 Dec;37(12):1341-4

4 Diosdado B, Stepniak DT, Monsuur AJ, **Franke L**, Wapenaar MC, Mearin ML, Koning F, Wijmenga C.
No genetic association of the human prolyl endopeptidase gene in the Dutch celiac disease population.
Am J Physiol Gastrointest Liver Physiol. 2005 Sep;289(3):G495-500

3 Diosdado B, Wapenaar MC, **Franke L**, Duran KJ, Goerres MJ, Hadithi M, Crusius JB, Meijer JW, Duggan DJ, Mulder CJ, Holstege FC, Wijmenga C.
A microarray screen for novel candidate genes in coeliac disease pathogenesis
Gut. 2004 Jul;53(7):944-51

2 **Franke L**, van Bakel H, Diosdado B, van Belzen M, Wapenaar M, Wijmenga C.
TEAM: a tool for the integration of expression, and linkage and association maps.
European J of Human Genetics. 2004 Aug;12(8):633-8

1 van Tilburg JH, Sandkuijl LA, **Franke L**, Strengman E, Pearson PL, van Haeften TW, Wijmenga C.
Genome-wide screen in obese pedigrees with type 2 diabetes mellitus from a defined Dutch population.
Eur J Clin Invest. 2003 Dec;33(12):1070-4

Colophon

Design	**Clever°Franke** Thomas Clever Gert Franke Lude Franke www.cleverfranke.com
Printer	Gildeprint Drukkerijen www.gildeprint.nl
Cover	80 gram g-print
Book Block	115 gram g-print
Colors	Pantone Black C, Pantone 8260 C, Pantone 805 C
Financial Support	The following financial sponsors are gratefully acknowledged: Celiac Disease Consortium Netherlands Genomics Initiative Netherlands Bioinformatics Centre Complex Genetics Group at the Department of Biomedical Genetics of the University Medical Centre Utrecht University Utrecht


50,000,000 bp
60,000,000 bp

70,000,000 bp

Deletion