Resequencing of positional candidates identifies low frequency *IL23R* coding variants protecting against inflammatory bowel disease.

Yukihide Momozawa[1], Myriam Mni[1], Kayo Nakamura[1], Wouter Coppieters[1], Sven Almer[2], Leila Amininejad[3], Isabelle Cleynen[4], Jean-Frédéric Colombel[5], Peter de Rijk[6], Olivier Dewit[7], Yigael Finkel[8], Miquel A. Gassull[9], Dirk Goossens[6], Debby Laukens[10], Marc Lémann[11], Cécile Libioulle[1], Colm O'Morain[12], Catherine Reenaers[13], Paul Rutgeerts[4], Curt Tysk[14], Diana Zelenika[15], Mark Lathrop[15], Jurgen Del-Favero[6], Jean-Pierre Hugot[16], Martine de Vos[10], Denis Franchimont[3], Severine Vermeire[4], Edouard Louis[13] & Michel Georges[1].

[1]Unit of Animal Genomics, GIGA-R & Faculty of Veterinary Medicine, University of Liège (B34), 1 Avenue de l'Hôpital, 4000-Liège, Belgium. [2]Division of Gastroenterology and Hepatology, IMK Linköpings Universitet, Linköping, Sweden. [3]Department of Gastroenterology, Erasme Hospital, ULB, 1070 Brussels, Belgium. [4]Department of Pathophysiology, Gastroenterology Section, Catholic University of Leuven, Leuven, Belgium. [5]Registre EPIMAD, Hôpital Calmette, Lille, France. [6]Applied Molecular Genomics, Department of Molecular Genetics, VIB, University of Antwerp, Universiteitsplein 1, 2610 Antwerp, Belgium. [7]Department of Gastroenterology, Clinique Universitaire St Luc, UCL, Brussels, Belgium. [8]Department of Gastroenterology, Karolinska Children's Hospital, Stockholm, Sweden. [9]Gastroenterology Department, Hospital Universitari Germans Trias i Pujol, 5ª planta, edifici general, Ctra del Canyet, s/n, 08916 Badalona, Spain. [10]Department Gastroenterology, University Hospital, Ghent University, 9000 Gent, Belgium. [11]Department of Gastroenterology, AP-HP, Hôpital Saint-Louis, Université Paris Diderot Paris-VII, Paris, France. [12]Adelaide and Meath Hospital, Dublin, Ireland. [13] Unit of Hepato-gastroenterology, GIGA-R & Faculty of Medicine, University of Liège (B34), 1 Avenue de l'Hôpital, 4000-Liège, Belgium. [14]Department of Gastroenterology, Örebro Medical Center Hospital, Örebro, Sweden. [15]Centre National de Génotypage, Evry, France. [16]INSERM U843, Hopital Robert Debré, 75019 Paris, France.

Correspondence: michel.georges@ulg.ac.be

Genome-wide association studies (GWAS) have identified tens of risk loci for many complex disorders including Crohn's disease (CD)[1,2]. However, common disease-associated SNPs explain at most ~20% of genetic variance. Several factors may account for the missing heritability[3-5], including rare risk variants not adequately tagged in GWAS[6-8]. That rare susceptibility variants indeed contribute to the variation of multifactorial phenotypes has been demonstrated for colorectal cancer[9], plasma levels of HDL cholesterol[10], blood pressure[11], type I diabetes[12], hypertriglyceridemia[13] and - in the case of CD – for the *NOD2* gene[14,15]. We herein describe the use of high-throughput resequencing of DNA pools to search for rare coding variants influencing susceptibility to CD in 63 GWAS-identified positional candidate genes. We report low frequency coding variants conferring protection against inflammatory bowel disease (IBD) in the *IL23R* gene, yet conclude that rare coding variants in positional candidates don't make a large contribution to inherited predisposition to CD.

A meta-analysis of three GWAS resulted in the identification of 30 significant and 10 suggestive susceptibility loci for CD[2]. The average confidence interval was 233 Kb (range: 20 to 1,140) encompassing 4.1 genes (range: 0 to 37) for a total of 153 positional candidates (Supplemental Table 1). We decided to sequence the open reading frame (ORF) and intron-exon boundaries of the 51 genes mapping to loci containing between one and five genes. For loci with more than six candidates, we retained 15 genes that mapped to significant networks identified when analyzing all candidates with Ingenuity Pathways (v8.5) (Supplemental Table 2). To these 66 genes, we added the *SLC22A4* candidate[16], as well as *PTGER4, ORMDL3* and *GSDMB* on the basis of reported cis-eQTL effects[2,17]. The list of 70 selected genes is provided in Supplemental Table 3.

After extensive optimization (cfr. Supplemental note 1), we selected a protocol involving (i) constitution of equimolar pools of genomic DNA from sets of 32 cases or controls, (ii) amplification, using Phusion Hot Start High Fidelity DNA Polymerase (Finnzymes Oy), of the 70 targeted ORFs and intron-exon boundaries as a series of 1,045 amplicons averaging 222 bp (range: 136-337 bp) (iii) equimolar pooling of ~300 amplicons, (iv) massive parallel pyrosequencing using the Roche FLX system[18] targeting an average sequence depth of 500 for both the Watson and Crick (W&C) strands, (v) detection of DNA sequence variants (DSV) using the Amplicon Variant Analyzer (AVA) software (Roche) augmented with custom-made scripts (Methods).

We opted for a staged design in which all 70 candidate genes would first (stage I) be sequenced on 112 cases and 112 controls. This provides 98.5% and 73.3% nominal power ($p \leq 0.05$) to detect the 12% and 7% excess of rare *NOD2* variants reported by Hugot *et al.*[14] and Lesage *et al.*[15] respectively. The most promising genes would be further evaluated on additional pools of cases and controls (stage II). To increase the impact of genetic effects other than *NOD2*, the 112 stage I cases did not carry either of three known *NOD2* susceptibility variants (*p.R702W*, *p.G908R* and *p.A1007fs*). To avoid subtle stratification, the corresponding 112 controls underwent the same selection. All analyzed cases and controls were of European decent.

92.9% of amplicons, corresponding to 63/70 genes (Supplemental Table 3) and 108.3 kb, could be sequenced with coverage ≥ 200 for both W&C strand in at least one case and one control pool. Simulations indicate that this coverage provides ≥ 83.4% power to detect singletons (i.e. one variant chromosome in DNA pool of 32 individuals) given the settings of the AVA software and self-imposed curation filters (Supplemental Fig.1). Average sequence depth (± SD) of retained amplicons was 1,471 ± 849 in cases and 1,420 ± 822 in controls (Supplemental Table 3). Analysis of the flowgrams yielded 372 DSV (Table 1, Supplemental Table 4). Transitions accounted for 82.5%, transversions for 16.1%, dinucleotide substitutions for 0.3% and indels for 1.1% of the variants. Synonymous (S) variants accounted for 41.7%, missense variants for 55.9%, nonsense variants for 0.8%, in-frame indels for 1.1%, and "boundary" (intronic within 2 bp from exon) (β) variants for 0.5%. DSV with estimated MAF < 0.05 amounted to 78.5%, while singletons represented 50.3% of the total (Fig. 1). As expected and reflecting purifying selection on (mildly) deleterious variants, the frequency spectrum of non-synonymous (NS) variants was shifted towards lower frequencies when compared to S variants. NS variants represented 60% of the total for MAF < 0.05 versus 40% for MAF ≥ 0.05 (Table 1 & Fig. 1). The high transition/transversion ratio (5.1) is thought to be due to (i) the analysis of ORF, as transversions are more likely to be NS and selected against, (ii) idiosyncrasies of the analyzed set of genes as their transition/transversion ratio tended to be higher than that of other ORF in HapMap data, and (iii) the elimination of low frequency (< 2.5%) C-A=G-T variants (Supplemental Table 5 and Supplemental note 1).

We evaluated our protocol in terms of sensitivity (fraction of true variants called), positive predictive value (PPV; fraction of true variants amongst called variants), and accuracy in estimating allelic frequency, focusing first on common variants (MAF ≥ 0.05). Analysis of the HapMap data revealed 62 bona fide SNPs with MAF ≥ 0.05 that were covered by the 879 retained amplicons. Five of these (= 8.1%) would lie within 6 bp from a homopolymer track and ignored (Methods). The remaining 57 were all detected, pointing towards excellent sensitivity. The 24 called SNPs with MAF ≥ 0.05 that were not genotyped in HapMap were inventoried in dbSNP (22/24 SNPs) or confirmed by the 1,000 Genomes Project (www.1000genomes.org/) (2 remaining), indicating excellent PPV. To evaluate the accuracy in estimating allelic frequencies we took advantage of 31 common SNPs that had been genotyped on the same individuals as part of other projects. Supplemental Fig. 2 shows the correlation between allelic frequency estimated from the genotyping data and read counts. The regression coefficient was 0.975 and the correlation 0.993.

To obtain similar estimates for rare variants, we manually (Sanger sequencing on ABI3730) sequenced 2,283bp of the *NOD2* ORF on the same 112 cases and 112 controls. Sanger sequencing revealed 38 variants with MAF < 0.05. Assuming faultless Sanger sequencing, sensitivity and PPV of the massive parallel resequencing protocol were 82.4% and 97.9%. Frequency estimates from read counts tended to underestimate actual frequencies of rare variants (regression coefficient: 0.822; correlation: 0.578) (Supplemental Fig. 2). We observed no difference in sequence depth between amplicon x DNA pool combinations in which rare DSV were detected and those in which no such DSV were found (Supplemental Fig. 3).

Having evaluated the performances of our protocol, we initiated the search for differences in cumulative frequencies of rare variants (MAF < 0.05) between cases and controls. Statistical significance of the observed differences was estimated on a gene-by-gene basis using a permutation test (Methods). P-values were computed for (i) S variants, (ii) all NS+β variants, and (iii) NS variants predicted by SIFT[19] to be damaging. For each gene x DSV-type combination we computed two p-values corresponding, respectively, to an enrichment of rare variants in cases (i.e. risk variants), or an enrichment of rare variants in controls (i.e. protective variants). Thus, we made the hypothesis that disruptive variants would increase disease-risk in some genes, while decreasing disease-risk in others. When applying a Bonferroni correction, none of the 63 sequenced genes showed a significant ($p <$ 7.94×10^{-4}) enrichment of rare variants neither in cases nor in controls, whether S, NS+β or damaging (Supplemental Table 6). There was no evidence for a difference in the distribution of p-values between S and NS+β variants, whether considering variants independently or on a gene-by-gene basis (Supplemental note 2). However, *NOD2* was showing the expected enrichment of rare NS variants (excluding the well known *p.R702W*, *p.G908R*, and *p.A1007fs* DSV) in cases (nominal $p = 5.94$ $\times 10^{-3}$; rank 3).

We therefore decided to pursue the sequencing (stage II) of the top 20% (i.e. 12) genes on 288 to 928 additional cases and 288 to 1,216 additional controls, depending on intermediate results. The procedure was identical to stage I: high-throughput resequencing of pooled amplicons obtained from DNA pools of cases or controls (32 individuals/pool; up to 29 case and 38 control pools/gene). Amplicons were appended with DNA pool-specific tags allowing simultaneous sequencing of multiple DNA pools. Average sequence depth in stage II was 988 ± 512 (range: 411 – 13,506) in cases, and 1,019 ± 415 (range: 405 – 10,414) in controls (Supplemental Fig. 1 & Supplemental Table 3). We detected 2 new common and 112 new rare variants (Supplemental Table 7). We observed no difference in sequence depth between amplicon x DNA pool combinations in which rare DSV were detected and those in which no such DSV were found (Supplemental Fig. 3).

We tested for differences in cumulative frequencies of rare variants in cases and controls using the same permutation test as above except that we only tested the significance of the enrichment with same polarity as in stage I, i.e. enrichment either for rare risk variants in cases (*FGFR1OP*, *GSDMB*, *IKZF3*, *IL1RL1*, *NOD2*, *SLC9A4*, *TNFSF8*) or rare protective variants in controls (*CCL8*, *CDKAL1*, *ENOX1*, *IL23R*, *SLC22A5*). At the outcome of stage II (Table 2), one gene yielded a suggestive association (*FGFR1OP*; nominal *p* = 0.040; Bonferonni-corrected *p* = 0.386), and one gene yielded a significant association (*IL23R*; nominal *p* = 2.67 x 10^{-3}; Bonferonni-corrected *p* = 0.0314).

Closer examination of *FGFR1OP* revealed that three NS variants (*p.T184I, p.K251N, p.S281P*) located within 2,436 bp from each other were segregating identically across DNA pools. This strongly suggested that they were in complete LD. When considering them as a single event, nominal p-values dropped to 0.124 in stage I and 0.081 in stage II. Hence, *FGFR1OP* was not considered for further analysis.

The *IL23R* signal was entirely due to three variants (*p.R86Q*, *p.G149R* and *p.V362I*) with cumulative frequency of 0.0052 in cases versus 0.0370 in controls in stage I and 0.0088 in cases versus 0.0193 in controls in stage II (Table 3 and Supplemental Table 8). The observation of an enrichment in controls of these presumably protective *IL23R* variants was consistent with the protective effect of *p.R381Q* that lead to the discovery of *IL23R* by GWAS[20]. *p.R381Q* was enriched in our controls as expected (*p* = 8.49 x 10^{-9}).

Being low frequency rather than very rare DSV[4] allowed targeted genotyping in independent samples. We developed TaqMan assays for *p.R86Q*, *p.G149R* and *p.V362I* in addition to *p.R381Q*. We first genotyped the sequenced individuals, which confirmed the enrichment of the *p.R86Q*, *p.G149R* and *p.V362I* variants in controls (Table 3 and Supplemental Fig. 2). We then genotyped an additional 1,565 CD patients, 2,000 controls and 3,101 familial samples (740 affected, 2,361 non-affected) (stage III). All analyzed individuals were of European decent and most of them previously used in GWAS replications. *p.G149R* (*p* = 0.022) and *p.V362I* (*p* = 1.51 x 10^{-3}) were significantly enriched in controls in the replication cohort while a similar trend, albeit not strictly significant, was observed for the rarer *p.R86Q* variant (*p* = 0.057) (Table 4).

p.R381Q confers protection against ulcerative colitis (UC)[20] (as well as ankylosing spondylitis[21]). We thus genotyped a cohort of 1,251 European decent UC patients for the same four *IL23R* DSV. Both *p.R381Q* (p = 9.03 x 10^{-9}) and *p.V362I* (p = 8.31 x 10^{-3}) were significantly depleted in UC patients, while the expected trend was observed for *p.G149R* (p = 0.087), but not for *p.R86Q* (p = 0.613) (Table 5).

We herein describe the systematic search for rare coding variants influencing inherited predisposition to CD in 63 positional candidates identified by GWAS. We report three novel low frequency *IL23R* variants protecting against CD: *p.R86Q*, *p.G149R* and *p.V362I*. The three same variants were found to be protective in an independent study, hence strengthening our claims (M. Rivas & M. Daly, personal communication). We present preliminary evidence that *p.G149R* and *p.V362I* act protectively against UC as well, as would be expected from the equivalent effect of *p.R381Q*.

As for the previously described *p.R381Q*, the novel *p.R86Q*, *p.G149R* and *p.V362I* variants are assumed to be hypomorphs dampening IL23R signaling. *p.G149R* and *p.R381Q* affect extremely conserved residues in the extracellular and intracellular domain of the receptor, respectively, and are predicted by SIFT[19] to be damaging. *p.R86Q* and *p.V362I*, on the contrary, affect poorly conserved residues and are predicted to be "tolerated" by SIFT[19] (using sequence information only) and "benign" by PolyPhen[22] (using sequence and structural information). Moreover, the reference *IL23R* sequences of some mammals carry the *Q* and *I* residues associated with IBD in human. While we cannot exclude that *p.R86Q* and *p.V362I* are enriched in cases because of their association with causative variants lying outside the ORF, we consider it more parsimonious that they affect IL23R signaling directly. Of note, relative protection conferred by the "damaging" *p.G149R* and *p.R381Q* (2.98 and 2.75) tended to be higher than that conferred by the "tolerated" *p.R86Q* and *p.V362I* (2.50 and 1.76), and the same tendency was observed for UC.

Relative protection against CD conferred by the newly detected low frequency variants was ~2.4 on average. Although possible overestimated (winner's curse), this value appears considerably larger than the ~1.2 relative risk conferred by the bulk of common effects detected in GWAS, and supports an increase in effect size with decreasing frequency[6]. However, the newly detected variants jointly explain only ~0.18% of the variance of the underlying liability, to be compared with ~0.85% for the more common *p.R381Q* and *rs7517847* variants[19] (Supplemental note 3). Haplotype analysis indicates that protection conferred by *p.R86Q, p.G149R* and *p.V362I* is largely independent of the more common *p.R381Q* and *rs7517847*, i.e. we provide no evidence for “synthetic association” [23] at the *IL23R* locus (Supplemental note 3).

Although not significant when accounting for multiple testing, we obtained evidence suggesting an enrichment of rare NS *NOD2* risk variants in cases in stage I, supporting previous reports[14,15]. This enrichment was not confirmed in stage II despite the sequencing of 928 cases and 992 controls. This discrepancy may be related to the selection of stage I samples carrying neither of the previously described *p.R702W*, *p.G908R* or *p.A1007fs NOD2* susceptibility variants, which were consequently enriched in stage II samples. Considering stage I and II samples jointly, however, indicates that the excess *NOD2* mutation load in CD cases is likely to be lower than previously assumed[14,15], more in line with recent estimates from a similarly conducted North-American scan for rare variants (M. Rivas & M. Daly, personal communication).

Our findings are highly reminiscent of those of Nejentsev et al.[12], who resequenced the ORF and regulatory regions of ten candidates for type I diabetes in 480 cases and 480 controls and reported four low frequency protective variants in *IFIH1*. These modest success rates contrast with Johansen et al.[13] who reported an enrichment of rare variants associated with hypertriglyceridemia (HTG; defined as fasting plasma triglyceride concentrations above the 95th percentile) in all four resequenced (438 cases and 327 controls) candidate genes from GWAS (*APOA5, GCKR, LPL, APOB*). We performed simulations indicating that this discrepancy is more likely to result from a difference in genomic architecture of the studied traits rather than from methodological ideosyncrasies (high-throughput sequencing of DNA pools in two stages versus Sanger sequencing of individual samples in one stage) (Supplemental note 4).

This study confirms the enrichment of low frequency variants (either in cases or controls) in at least some genes underlying inherited predisposition to complex diseases. Our results support an increase in effect size with decreasing variant frequency. However, because of their frequency rare variants explain less of the heritability than their common counterparts. Achieving adequate power to reliably detect low frequency variants will require resequencing of cohorts larger than in this study. This will become increasingly feasible as sequencing technology continues to improve. The demonstration of an enrichment of rare or low frequency variants in candidate genes could then become an effective way to demonstrate the causality of candidate genes from GWAS.

Acknowledgments

This work was funded by the Walloon Region (IPSEQ, Crohn & CIBLES projects), by the Politique Scientifique Fédérale (IAP GENFUNC), by the Fonds National de la Recherche Scientifique (FNRS), by the Communauté Française de Belgique (ARC BIOMOD). EL is a senior research associate from the FNRS. KN benefitted from a post-doctoral fellowship from the University of Liège. We are grateful to Alex Kvasz, Frédéric Farnir and Denis Baurain for their help with bioinformatic analyses. We thank the personnel of the Flanders Institute for Biotechnology (VIB) Genetic Service Facility (www.vibgeneticservicefacility.be) and the GIGA-R genotranscriptomic platform for their contribution to sequencing.

AUTHOR CONTRIBUTIONS

YM, MM, KN, LA, DG, DZ performed experiments; YM, WC, PdR, MG analyzed data; MLa, JDF supervized experiments; SA, LA, JFC, OD, YF, MAG, MLé, CO, CR, PR, CT, JPH, MdV, DF, SV, EL examined patients and collected samples; IC, DL, CL prepared and organized samples; YM and MG wrote the manuscript.

Figure 1: Frequency distribution of MAF for (i) synonymous (blue), (ii) all non-synonymous (red), (iii) damaging (SIFT[19]) non-synonymous DSV (orange).

Methods

High-throughput pyrosequencing on Roche FLX. Genomic DNA concentrations were determined by Quant-iT PicoGreen dsDNA Reagent and Kits (Invitrogen) for the constitution of equimolar pools of 32 cases or controls (except one pool of 48 in stage I). Primer pairs for PCR were selected using Primer 3[24], avoiding known SNP positions. Amplicon-specific PCR reactions were set up in 30 µl volumes containg 6 µl of 5x Phusion HF buffer, 200 µM of each dNTP, 0.5 µM of each primer, and 0.6 U of Phusion High-Fidelity DNA Polymerase (Finnzymes Oy). Cycling conditions were 98 °C for 2 min, 32 cycles at 98 °C for 10 sec, 60 °C for 30 sec and 72 °C for 15 sec, followed by 72 °C for 10 min on a GeneAmp PCR System 2700 thermal cycler (Applied Biosystems). PCR products were purified using MultiScreen $\text{PCR}_{\mu 96}$ Filter Plates (Millipore), and quantified with the Quant-iT PicoGreen dsDNA reagent and kit. Up to 300 amplicons were combined in equimolar ratios. Pooled amplicons were concentrated using the Montage PCR Filter Units (Millipore) and purified using the AMPure kit (Agencourt Biosciences). Final concentration and length distribution were measured using the Experion DNA 1K Analysis kit (Bio-Rad) on Experion Automated Electrophoresis Station (Bio-Rad). High-throughput pyrosequencing was carried out using both primer A and B on a Roche 454 Genome Sequencer FLX instrument following the recommendations of the manufacturer[18].

DSV detection. Image and data were processed with the Genome Sequencer FLX System Software Package (Roche). DSV were extracted from sff files using the AVA software. AVA reports DSV if observed at least four times and representing ≥ 0.5% of the reads. From the AVA-generated list, we eliminated DSV (i) unless observed on both W&C strand, (ii) with flanking DSV within 2-bp on both sides, (iii) in or within 6-bp from a homopolymer (≥5x) track, (iv) corresponding to C/A or G/T substitutions with frequency < 0.025 (cfr. Supplemental note 1).

Testing for a differential load of rare variants in cases and controls from resequencing data. Excess load of rare variants in cases (risk variants) or controls (protective variants) was tested on a gene-by-gene basis, and - within gene – by DSV variant type (S, NS+β, damaging). Rare variants were defined as DSV with MAF < 0.05. The results were essentially unaffected by the threshold frequency used to define rare variants (0.02 - 0.05) (data not shown). DSV read counts (number of reads with the DSV/total number of reads for that amplicon) were converted to the closest chromosome counts (>0) (number of chromosomes with the DSV/total number of chromosomes in the pool = 64) and these were summed over DNA pools, separately for cases and controls. The p-value of the difference in DSV chromosome counts between cases and controls was then computed using two one-tailed Fisher's exact tests, one testing an excess in cases (risk), the other in controls (protective). For a given gene, we then multiplied hypothesis-specific (risk and protective) p-values across rare variants, to generate two gene-specific summary p-values. The statistical significance of these summary statistics was estimated by permutation testing. For each rare DSV, case vs control status of mutant chromosomes were assigned randomly yet accounting for the possibility that the number of successfully sequenced chromosomes (i.e. DNA pools) might differ between cases and controls. The same two gene-specific summary p-values were generated for 1,000,000 permutations, and the significance of the p-values obtained with the real data estimated as the proportion of permutations with lower, hypothesis-specific, summary p-value.

Case-control and familial association test based on individual genotypes. SNPs were tested on individual DNA using custom TaqMan assays (Applied Biosystems). The statistical significance of the difference in DSV frequency between cases and controls was estimated using one-side Fisher's exact test. The familial cohort was used to evaluate the significance of the distorted segregation of DSV *p.R381Q* and *p.V362I* (TDT) from heterozygous parents to affected offspring using a custom-made script. As no heterozygous parents were available in the familial cohort for DSV *p.R86Q* and *p.G149R,* one affected individual per family was added to the case cohort in the case-control analysis for the analysis of these variants. Combining test statistics were done across resequencing, case-control and TDT experiments using a permutation test akin to the one described above.

References

1. Altshuler, D., Daly, M.J. & Lander, E.S. Genetic mapping in human disease. *Science* **322**, 881-8 (2008).
2. Barrett, J.C. et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* **40**, 955-62 (2008).
3. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-9 (2010).
4. Manolio, T.A. et al. Finding the missing heritability of complex diseases. *Nature* **461**, 747-53 (2009).
5. Kong, A. et al. Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868-74 (2009).
6. Pritchard, J.K. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* **69**, 124-37 (2001).
7. Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* **40**, 695-701 (2008).
8. Kruglyak, L. The road to genome-wide association studies. *Nat Rev Genet* **9**, 314-8 (2008).
9. Fearnhead, N.S. et al. Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. *Proc Natl Acad Sci U S A* **101**, 15992-7 (2004).
10. Cohen, J.C. et al. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**, 869-72 (2004).
11. Ji, W. et al. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet* **40**, 592-9 (2008).
12. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J.A. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324**, 387-9 (2009).
13. Johansen, C.T. et al. Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat Genet* **42**, 684-7 (2010).

14. Hugot, J.P. et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599-603 (2001).

15. Lesage, S. et al. CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *Am J Hum Genet* **70**, 845-57 (2002).

16. Peltekova, V.D. et al. Functional variants of OCTN cation transporter genes are associated with Crohn disease. *Nat Genet* **36**, 471-5 (2004).

17. Libioulle, C. et al. Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet* **3**, e58 (2007).

18. Margulies, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-80 (2005).

19. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073-81 (2009).

20. Duerr, R.H. et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* **314**, 1461-3 (2006).

21. Burton, P.R. et al. Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nat Genet* **39**, 1329-37 (2007).

22. Ramensky, V., Bork, P. & Sunyaev, S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* **30**, 3894-900 (2002).

23. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D.B. Rare variants create synthetic genome-wide associations. *PLoS Biol* **8**, e1000294 (2010).

24. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**, 365-86 (2000).