LETTER

doi:10.1038/nature12170

Negligible impact of rare autoimmune-locus coding-region variants on missing heritability

Karen A. Hunt[1], Vanisha Mistry[1], Nicholas A. Bockett[1], Tariq Ahmad[2], Maria Ban[3], Jonathan N. Barker[4], Jeffrey C. Barrett[5], Hannah Blackburn[5], Oliver Brand[6], Oliver Burren[7], Francesca Capon[4], Alastair Compston[3], Stephen C. L. Gough[6], Luke Jostins[8], Yong Kong[9], James C. Lee[10], Monkol Lek[11], Daniel G. MacArthur[11], John C. Mansfield[12], Christopher G. Mathew[4], Charles A. Mein[13], Muddassar Mirza[4], Sarah Nutland[7], Suna Onengut-Gumuscu[14], Efterpi Papouli[4], Miles Parkes[10], Stephen S. Rich[14], Steven Sawcer[3], Jack Satsangi[15], Matthew J. Simmonds[6], Richard C. Trembath[16], Neil M. Walker[7], Eva Wozniak[13], John A. Todd[7], Michael A. Simpson[4], Vincent Plagnol[17] & David A. van Heel[1]

Genome-wide association studies (GWAS) have identified common variants of modest-effect size at hundreds of loci for common autoimmune diseases; however, a substantial fraction of heritability remains unexplained, to which rare variants may contribute[1,2]. To discover rare variants and test them for association with a phenotype, most studies re-sequence a small initial sample size and then genotype the discovered variants in a larger sample set[3–5]. This approach fails to analyse a large fraction of the rare variants present in the entire sample set. Here we perform simultaneous amplicon-sequencing-based variant discovery and genotyping for coding exons of 25 GWAS risk genes in 41,911 UK residents of white European origin, comprising 24,892 subjects with six autoimmune disease phenotypes and 17,019 controls, and show that rare coding-region variants at known loci have a negligible role in common autoimmune disease susceptibility. These results do not support the rare-variant synthetic genome-wide-association hypothesis[6] (in which unobserved rare causal variants lead to association detected at common tag variants). Many known autoimmune disease risk loci contain multiple, independently associated, common and low-frequency variants, and so genes at these loci are a priori stronger candidates for harbouring rare coding-region variants than other genes. Our data indicate that the missing heritability for common autoimmune diseases may not be attributable to the rare coding-region variant portion of the allelic spectrum, but perhaps, as others have proposed, may be a result of many common-variant loci of weak effect[7–10].

Recent large-scale human sequencing studies have revealed an abundance of rare variants (which we define as minor allele frequency (MAF) $< 0.5\%$) and shown that these are geographically localized and are more likely to have deleterious functional consequences[11,12]. In the largest sample size studied to date[12], 202 genes in 14,002 people were re-sequenced, and ~95% of exonic variants identified were found to be rare, with 74% observed in only one or two subjects. More broadly, across ~15,000 genes, similar findings were observed in recent exome-sequencing studies of 2,440 and 6,515 subjects[13,14]. Importantly, these studies demonstrate that even if we had reference variation databases from a million subjects, most of the rare-variant allelic spectrum of any given sample set (for example, a case–control cohort) will be unique and only identifiable by direct re-sequencing of the entire sample set.

There are only a handful of published examples of rare coding-region variants associated with common autoimmune diseases (although many examples in familial/Mendelian immune-mediated diseases). Coding-region variants in *IFIH1* associated with type 1 diabetes (MAF in controls = 0.67–2.2%)[3], *TYK2* with multiple autoimmune diseases[15] and *IL23R* with inflammatory bowel disease[5], for example, are low frequency (which we define as MAF = 0.5–5%) rather than particularly rare. In other examples, the existing evidence for association, and/or the effect sizes, are relatively weak (for example, *CARD14* and psoriasis[16], *IL2RA* and *IL2RB* and rheumatoid arthritis[17]). The association of rare coding-region variants of *NOD2* (also known as *CARD15*) in Crohn's disease probably provides the best example, albeit three low-frequency variants comprise over 80% of all the disease-causing mutations[18]. Most of the studies also lose power (especially for tests in which multiple rare variants are pooled into a single analysis, for example by gene) by initially sequencing only a small sample subset rather than testing the entire rare-variant content of a large case–control sample set. We sought to improve on these methods by performing highly multiplexed sequencing of sufficiently high quality to enable direct genotyping in the entirety of a large autoimmune disease case–control collection.

We selected subjects from a single population—individuals of white Northern-European ethnicity living in the UK (Methods)—to minimize any effects of population stratification. We selected to re-sequence all RefSeq exons for 25 genes from 20 GWAS-identified risk loci showing overlap between six common autoimmune disease phenotypes (autoimmune thyroid disease, coeliac disease, Crohn's disease, psoriasis, multiple sclerosis and type 1 diabetes). All genes studied were from risk loci for at least two phenotypes, all genes had known immune system function, 18 out of 20 loci had either a single candidate immune gene or all immune genes at a locus were selected (the remaining two loci had partial transcripts of another immune gene within the 0.1 centimorgan (cM) linkage disequilibrium block), and all genes and loci were densely genotyped on the Illumina ImmunoChip (Supplementary Table 1)[19]. We attempted high-throughput sequencing of 52,224 samples (including positive and negative controls, and repeats). We performed extensive quality control on both samples and variant calls (Methods). The final data set comprised 41,911 phenotyped individuals (autoimmune disease cases and controls), with ImmunoChip

[1]Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London E1 2AT, UK. [2]Peninsula College of Medicine and Dentistry, Barrack Road, Exeter EX2 5DW, UK. [3]University of Cambridge, Department of Clinical Neurosciences, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK. [4]Division of Genetics and Molecular Medicine, King's College London School of Medicine, 8th Floor Tower Wing, Guy's Hospital, London SE1 9RT, UK. [5]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. [6]Oxford Centre for Diabetes Endocrinology and Metabolism, University of Oxford, Oxford OX3 7LJ, UK. [7]Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Cambridge CB2 0XY, UK. [8]Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK. [9]Department of Molecular Biophysics and Biochemistry, W.M. Keck Foundation Biotechnology Resource Laboratory, Yale University, New Haven, Connecticut 06510, USA. [10]Department of Medicine, University of Cambridge School of Clinical Medicine, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK. [11]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [12]Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne NE1 3BZ, UK. [13]Genome Centre, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, John Vane Science Centre, Charterhouse Square, London EC1M 6BQ, UK. [14]Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia 22908-0717, USA. [15]Gastrointestinal Unit, Molecular Medicine Centre, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK. [16]Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London E1 2AT, UK. [17]University College London Genetics Institute, Gower Street, London WC1E 6BT, UK.

Table 1 | Variant types in protein-coding regions of 25 genes in 41,911 phenotyped individuals

Variant type	All variants	Rare (MAF < 0.5%)*	Novel†
Nonsynonymous SNV	1,792	1,758	1,379
Splicing SNV	86	85	65
Stopgain SNV	47	47	42
Synonymous SNV	1,024	972	674
Frameshift indels	31	31	31
Nonframeshift indels	10	10	10
Total variants	2,990	2,903	2,201
Singleton	1,602	1,598	1,411
Doubleton	470	468	378

Numbers shown are after quality-control steps. Annotation performed with GENCODE V14 gene definitions. Triallelic ($n = 124$) and quadrallelic ($n = 3$) sites (combined SNVs and indels) are shown as multiple separate variants with the appropriate annotation for each non-reference allele.
* MAF in 17,019 sequenced controls.
† Not seen in dbSNP137, or 1000 Genomes Project (April 2012 release), or NHLBI (data release ESP6500SI, with 6,503 individuals).

array genotypes available for 32,806 of these individuals (Supplementary Table 2). We discovered 4,377 variant sites across all amplicons, and the genotype call rate was 99.9989% (reference homozygote as well as non-reference genotypes) across 41,911 individuals. Of these, 2,990 variants were in protein-coding regions (including exon splice sites) of the 25 genes (Table 1 and Supplementary Table 3); 97.1% of which are rare (MAF in 17,019 controls, <0.5%); 73.6% are novel when compared with current published data sets (dbSNP137, 1000 Genomes Project, National Heart, Lung, and Blood Institute (NHLBI)) containing >6,000 individuals and 67.3% are novel compared to an unpublished data set of 25,994 exome-sequenced individuals (D. G. MacArthur, personal communication); and 68.9% were only seen in one (singleton) or two (doubleton) individuals. These proportions of novel, and rare, variants are similar to recent data from other large re-sequencing studies[12].

Our very high coverage data (99.8% of 183.4 million (site X sample) genotype calls had a read depth of ≥40 and 96.6% had a read depth of >100; Supplementary Fig. 1) enabled stringent data filtering on call rate per sample, per variant site, and other criteria (Methods). To confirm data quality, we performed further experiments and analyses as follows: (1) we genotyped one control sample 296 times (on different 48-sample microfluidic chips), and the genotype call error rate was two non-consensus genotype calls of 1,295,581 called genotypes (0.00015%); (2) 32,806 out of 41,911 subjects also had dense ImmunoChip genotyping data at the 25 genes, and genotype concordance at 91 variant sites genotyped on both platforms was 99.994%; (3) transition/transversion (Ti/Tv) rates, a quality-control measure based on expected human mutation types, were 2.434 at coding-region variants (2.427 at singletons), 2.44 at rare (MAF < 0.5%) variants (2.437 at singletons) and 2.275 at novel variants (2.273 at singletons) (definitions in Table 1); (4) we selected all (35) nonsense single nucleotide variants (SNVs) and all (39) frameshift insertions/deletions (indels) in the ImmunoChip-genotyped samples for Sanger sequencing: two variants failed assay/PCR (polymerase chain reaction) design and there was one false-positive SNV and one false-positive indel (overall false-positive rate = 2.8%). All 70 validated SNVs and indels had the same alleles in high-throughput and Sanger-sequencing

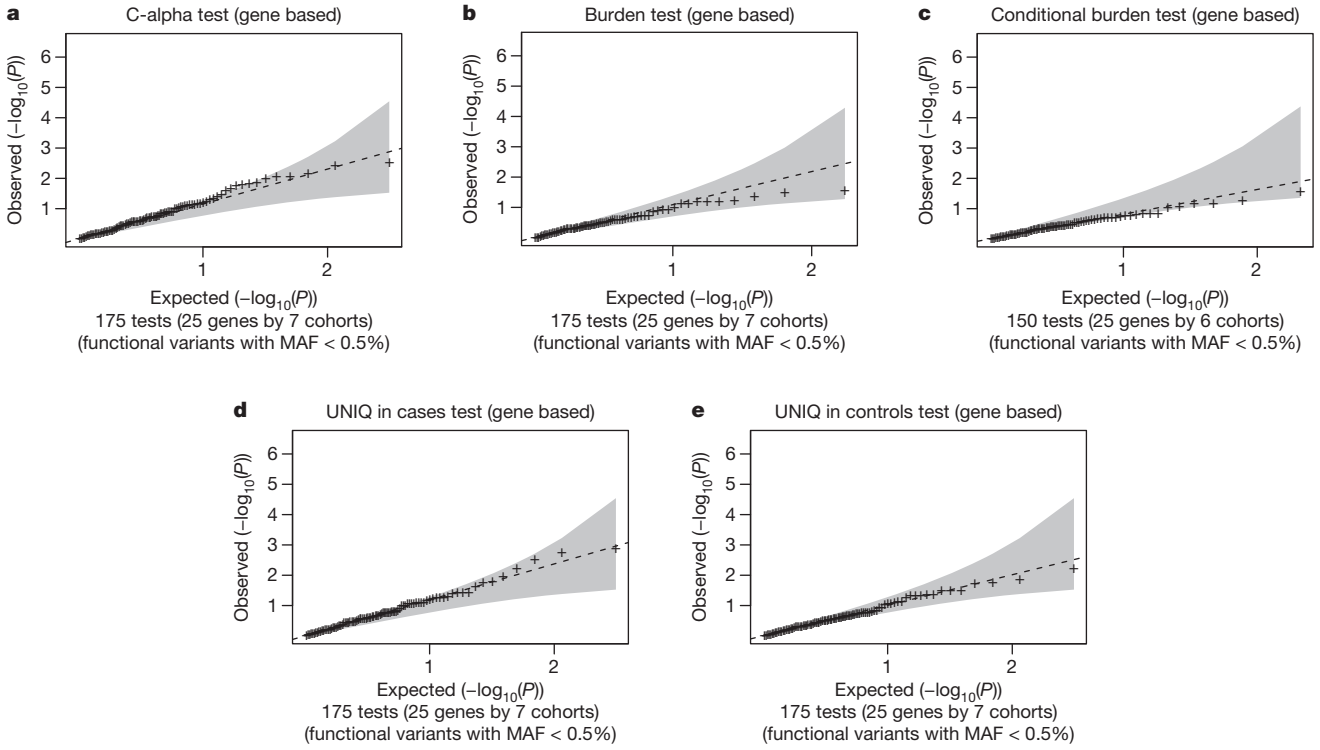

Figure 1 | Association analyses of discovered rare functional variants in autoimmune diseases. We define rare functional variants as MAF < 0.5% in 17,019 controls and predicted nonsynonymous, premature-stop or splice-site annotation. Quantile–quantile plots compare observed versus expected test-statistic distributions, with shading indicating 99% confidence intervals. Full results are available in Supplementary Data. Each of six individual diseases, and all autoimmune diseases combined, were tested as phenotypes. **a**, Gene-based C-alpha test (25 genes by 7 phenotypes, $n = 41,911$ subjects) allowing for both risk and protective effects for rare functional variants. Singleton variants pooled into a single binomial count per phenotype. **b**, Gene-based burden tests (25 genes by 7 phenotypes, $n = 41,911$ subjects) comparing summed allele counts for rare functional variants in cases versus controls with Fisher's exact test. **c**, Conditional gene-based burden test (25 genes by 6 phenotypes, $n = 32,806$ subjects): rare functional-variant allele counts are summed for each individual per gene and introduced in a logistic regression, including ImmunoChip covariates for multiple independent top (common) variant signals selected on the basis of a stepwise regression (down to $P > 10^{-4}$). The psoriasis phenotype was not tested as most samples do not have ImmunoChip data. **d**, Count of case-unique rare alleles (UNIQ) tests (25 genes by 7 phenotypes, $n = 41,911$ subjects): compares the number of rare functional variants only observed in cases with the distribution of this value upon random permutation (10,000 times) of the phenotypes. **e**, Count of control-unique rare alleles (UNIQ) tests: same as **d** but for rare functional variants uniquely observed in controls.

assays; (5) proportions of rare, and of known, variants were similar to those found by other large sequencing studies, and we identified no common or low-frequency novel variant sites.

We first attempted to identify any low-frequency or rare variants of larger effect. We performed for each coding-region variant and each of seven phenotypes (including all autoimmune disease cases combined) a single-variant association analysis. Only previously reported loci were observed with common variants (MAF > 5%), as expected. We identified three low-frequency (MAF = 0.5–5%) and rare (MAF in 17,019 controls = <0.5%) exonic variants with single SNP association $P < 10^{-4}$ (chosen as a partial Bonferroni multiple testing correction for 25 genes and 7 phenotypes, but not correcting for all variants per gene) (Supplementary Table 4 and Supplementary Data). We next analysed low-frequency and rare exonic variants, conditioning on common-variant non-coding signals at each locus, and observed no additional association signals (Supplementary Data). An association between type 1 diabetes and the low-frequency *UBASH3A* SNP rs17114930 was observed, but conditional regression analysis showed this signal to be secondary to a stronger common-frequency variant/haplotype previously identified by GWAS[20]. We identified novel low-frequency (nearly 'common' as MAF in 17,019 controls = 4.97%) *NCF2* coding-region variant associations with coeliac disease at two SNPs (rs17849502, nonsynonymous; rs17849501, synonymous; in almost complete linkage disequilibrium $r^2 = 0.992$). Both variants were present on the Illumina ImmunoChip, but just failed quality-control criteria in our previous coeliac disease study owing to missing data[19]. We replicated the UK findings in 4,313 coeliac cases and 3,954 controls (European samples, Methods; rs17849502 $P = 4.46 \times 10^{-5}$ (Cochran–Mantel–Haenszel test), odds ratio 1.35 (95% CI = 1.17–1.55)). Logistic regression analysis conditioning on rs17849502 in the UK re-sequencing data set revealed no further single-variant coeliac disease association signals below $P < 10^{-4}$. *NCF2* is a component of the neutrophil NADPH oxidase respiratory burst complex. Different disease-causing mutations cause the recessive Mendelian phenotype chronic granulomatous disease. The rs17849502/H389Q variant is also associated with the autoimmune disease systemic lupus erythematosus[21]. Functional studies have shown that the minor allele of rs17849502/H389Q reduces the binding efficiency of NCF2 to the guanine nucleotide-exchange factor VAV1 (ref. 21). These data now implicate a disease mechanism of impaired neutrophil function in coeliac disease, a condition previously thought to be of predominantly B- and T-cell-mediated immunopathogenesis, and where neutrophils may have a role in regulating adaptive immunity[22].

We noted that even with ~7,000 cases and ~17,000 controls the power to detect association signals using single-variant tests for variants (MAF < 0.5%) of modest effect (for example, odds ratio < 3) is limited (Supplementary Fig. 2) and therefore we performed gene-based pooled-variant association tests to better detect the combined effect of multiple variants. We defined coding-region variants as functional candidates if the variants were rare (MAF in 17,019 controls = <0.5%) and predicted to be of potential functional impact (nonsynonymous, premature stop, splice-site altering; see Methods). We pooled variants (by gene) in analyses to detect different scenarios (Fig. 1 and Supplementary Data), including the C-alpha test, which can detect a combination of risk and protective variants; burden tests to detect either an excess of risk variants in cases or protective variants in controls; a modified version of the burden test using conditional regression and common-variant non-coding signals at a locus as covariates; a test to detect an excess of rare variants seen uniquely in cases (the case or control unique tests being particularly suitable for the study of the large numbers of singleton and doubleton variants we observe); and a test to detect an excess of rare variants seen uniquely in controls. The distribution of association statistics for all five pooled gene tests across each of the six or seven phenotypes tested was consistent with the global null of no association.

On the basis of these results, in the largest (to the best of our knowledge) human disease sample sequencing study to date, we find little support for a significant impact of rare coding-region variants in known risk genes for the autoimmune disease phenotypes tested. Our data provide little stimulus in support of large-scale whole-exome sequencing projects in common autoimmune diseases. Using average genetic-effect estimates from our data (Methods), over all loci and phenotypes we have tested, we estimate that rare variants contribute to less than 3% of the heritability explained by common variants at these known risk loci[23].

METHODS SUMMARY

Sequencing. DNA (corresponding to exonic sequence of 25 autoimmune disease risk genes) was PCR-amplified in a multiplexed microfluidics assay (Fluidigm Access Array). PCR amplicons from a sample were pooled, and barcoded with one of 1,536 unique ten-base-pair sequences. Libraries of 1,536 samples were sequenced on Illumina HiSeq instruments. Reads were aligned to the GRCh37 human reference and SNVs and small indels called. Samples and called variants were extensively filtered on the basis of call rate and other criteria. Selected variants were validated by Sanger dideoxy sequencing. Genotype data from Illumina ImmunoChip array-based genotyping was merged with Fluidigm sequencing-based genotypes.

Statistical analysis. Statistical analysis was performed in R, and using PLINK/SEQ software.

Full Methods and any associated references are available in the online version of the paper.

Received 27 February; accepted 8 April 2013.
Published online 22 May 2013.

1. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).
2. Gibson, G. Rare and common variants: twenty arguments. *Nature Rev. Genet.* **13,** 135–145 (2012).
3. 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,** 387–389 (2009).
4. Rivas, M. A. *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nature Genet.* **43,** 1066–1073 (2011).
5. Momozawa, Y. *et al.* Resequencing of positional candidates identifies low frequency *IL23R* coding variants protecting against inflammatory bowel disease. *Nature Genet.* **43,** 43–47 (2011).
6. Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D. B. Rare variants create synthetic genome-wide associations. *PLoS Biol.* **8,** e1000294 (2010).
7. Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T. L. & Kruglyak, L. Finding the sources of missing heritability in a yeast cross. *Nature* **494,** 234–237 (2013).
8. Stahl, E. A. *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nature Genet.* **44,** 483–489 (2012).
9. Park, J. H. *et al.* Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genet.* **42,** 570–575 (2010).
10. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genet.* **42,** 565–569 (2010).
11. Coventry, A. *et al.* Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature Commun.* **1,** 131 (2010).
12. Nelson, M. R. *et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337,** 100–104 (2012).
13. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337,** 64–69 (2012).
14. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493,** 216–220 (2013).
15. Strange, A. *et al.* A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between *HLA-C* and *ERAP1*. *Nature Genet.* **42,** 985–990 (2010).
16. Jordan, C. T. *et al.* Rare and common variants in *CARD14*, encoding an epidermal regulator of NF-kappaB, in psoriasis. *Am. J. Hum. Genet.* **90,** 796–808 (2012).
17. Diogo, D. *et al.* Rare, low-frequency, and common variants in the protein-coding sequence of biological candidate genes from GWASs contribute to risk of rheumatoid arthritis. *Am. J. Hum. Genet.* **92,** 15–27 (2013).
18. Lesage, S. *et al. CARD15/NOD2* mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *Am. J. Hum. Genet.* **70,** 845–857 (2002).
19. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nature Genet.* **43,** 1193–1201 (2011).
20. Barrett, J. C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature Genet.* **41,** 703–707 (2009).
21. Jacob, C. O. *et al.* Lupus-associated causal mutation in neutrophil cytosolic factor 2 (NCF2) brings unique insights to the structure and function of NADPH oxidase. *Proc. Natl Acad. Sci. USA* **109,** E59–E67 (2012).
22. Kolaczkowska, E. & Kubes, P. Neutrophil recruitment and function in health and inflammation. *Nature Rev. Immunol.* **13,** 159–175 (2013).

23. Liu, D. J. & Leal, S. M. Estimating genetic effects and quantifying missing heritability explained by identified rare-variant associations. *Am. J. Hum. Genet.* **91,** 585–596 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements The study was primarily funded by the Medical Research Council (MRC G1001158 to D.A.v.H and V.P.), with further funding from Coeliac UK (to D.A.v.H). We thank C. Wijmenga and G. Trynka for sharing ImmunoChip data, and the International Multiple Sclerosis Genomics Consortium for ImmunoChip data and samples. J.N.B. and R.C.T. are supported by MRC grant G0601387. This research was supported by the National Institutes for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. The study was supported by the Cambridge NIHR Biomedical Research Centre. We thank E. Gray and D. Jones (Wellcome Trust Sanger Institute) for sample preparation. We acknowledge use of DNA from The UK Blood Services collection of Common Controls (UKBS-CC collection), funded by the Wellcome Trust grant 076113/C/04/Z and by NIHR programme grant to NHS Blood and Transplant (RP-PG-0310-1002). The collection was established as part of the Wellcome Trust Case Control Consortium (WTCCC). We acknowledge use of DNA from the British 1958 Birth Cohort collection, funded by the UK MRC grant G0000934 and the Wellcome Trust grant 068545/Z/02. We thank nurses and doctors for recruiting autoimmune thyroid disease (AITD) subjects into the AITD National Collection, funded by the Wellcome Trust grant 068181. We acknowledge use of DNA from the Cambridge BioResource. We acknowledge use of DNA from the Juvenile Diabetes Research Foundation (JDRF)/Wellcome Trust Case-Series (GRID), funded by JDRF and the Wellcome Trust (grant references JDRF 4-2001-1008 and WT061858). The subjects were recruited in the UK by D. Dunger and his team with support from the British Society for Paediatric Endocrinology and Diabetes. The samples were prepared and provided by the JDRF/Wellcome Trust Diabetes and Inflammation Laboratory, University of Cambridge, UK. Psoriasis samples used were based on the WTCCC2 GWAS clinical panel, for which we thank D. Burden, C. Griffiths, M. Cork and R. McManus. Finally, we would like to thank all autoimmune disease and control subjects for participating in this study.

Author Contributions D.A.v.H. designed and led the study. K.A.H. coordinated wet laboratory work, with K.A.H., V.M., N.A.B. and E.W. performing DNA sample preparation, Fluidigm PCR amplification, sample barcoding, MiSeq library validation and Sanger sequencing preparation. HiSeq sequencing was performed by M.M. and E.P. D.A.v.H., V.P. and M.S. performed bioinformatics and statistical analyses. All other authors contributed to diverse aspects of sample collection, phenotyping, DNA preparation, ImmunoChip data production or specific analyses. D.A.v.H. and V.P drafted the manuscript, which all authors reviewed.

Author Information Genome data has been deposited at the European Genome–phenome Archive (http://www.ebi.ac.uk/ega/), which is hosted at the EBI, under accession number EGAS00001000476. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.A.v.H. (d.vanheel@qmul.ac.uk) or V.P. (v.plagnol@ucl.ac.uk).

METHODS

Gene selection. All genes studied (listed in Supplementary Table 3) were risk loci for at least two phenotypes, had a known immune system function, were from loci with only a single strong candidate immune gene (or all immune genes were selected at four loci: *IL18R1*, *IL18RAP*; *CTLA4*, *CD28*, *ICOS*; *IL2*, *IL21*; *PTPRK*, *THEMIS*), and all genes and loci were densely genotyped with all 1000 Genomes pilot project variants on the Illumina ImmunoChip (for design of this chip, see ref. 19). Additional criteria favouring locus selection were: known multiple independent association signals, risk (not necessarily same variants/haplotype or signal direction) for many autoimmune diseases, fine-mapping or other data strongly suggesting a single candidate gene, and smaller complementary DNA size.

Samples. UK samples for the six component immune disease phenotypes have been described in previous publications (which also contain full details of Ethics Committee approvals)[19,20,24-27], as have the three control populations[19,28]. Informed consent was obtained from all subjects. Individuals with self-reported autoimmune disease were excluded from the UK Blood Services — Common Controls and NIHR Cambridge Biomedical Research Centre Cambridge BioResource controls. Samples with self-stated non-white European ethnicity were excluded (later further confirmed by ImmunoChip-based principal component ethnicity analysis for 32,806 samples). Samples with gross discordance with ImmunoChip genotypes and/or with known gender or genotype-mismatch issues from previous GWAS were excluded. Samples with known duplicates or relatedness (as distant as first cousins) were excluded, relatedness was later confirmed by ImmunoChip genome-wide identity-by-state analysis and by analysis of multiple rare-variant sharing in Fluidigm sequencing data. Additional independent European samples genotyped for rs17849502 (4,313 coeliac cases and 3,954 controls) were previously described[19].

Wet-lab. PCR primers were designed for all RefSeq exons of 26 genes, and amplicons selected to be 150–200 base pairs (bp) in size. There was minor primer design dropout at *IL18R1*, *STAT4*, *THEMIS* and *ZMIZ1*, although >94% of exon sequence was still covered at these genes. Variant calls at the gene *YDJC* later proved unreliable with highly biased allele depths at heterozygote sites, probably due to the very high exon GC content (~70%), and this gene was not further analysed nor is it discussed elsewhere in this study. The total length of (overlapping) amplicons was 95,927 bp; with primers removed (still overlapping) 72,612 bp; and with primers removed and unique sequence 58,550 bp. PCR amplification was performed using 50 ng genomic DNA per sample on the 48 sample/plate Fluidigm microfluidic Access Array system. PCR primers for 511 PCR reactions were pooled up to 12-plex per well in 48 pools. Individual per sample per pool PCR reactions took place in ~35-nl reaction chambers with ~300 DNA haplocopies per reaction. All pools per sample were combined. Each sample's pool was then individually barcoded in a second PCR reaction with one of 1,536 10-bp Fluidigm-designed unique barcodes (Fluidigm unidirectional sequencing protocol).

Sequencing. Thirty-four libraries (each of 1,536 barcoded samples) were generated. Libraries were first sequenced on an Illumina MiSeq for rapid quality control of the barcoding step, and to optimize loading concentrations/cluster density. Libraries were then sequenced one per lane using 101-bp paired-end reads and an 11-bp index read (the last base of each read being only used for chemistry cycle phasing purposes) on Illumina HiSeq sequencers. Lanes were repeated if target cluster density or target clusters passing filter were not achieved. Individual samples were de-multiplexed by Illumina CASAVA software, allowing zero mismatches per 10-bp barcode. Sanger sequencing was performed on PCR products using an ABI 3730xl DNA analyser and ABI big dye terminator 3.1 cycle chemistry. We sequenced all samples with rare-variant allele genotypes, and a control sample, for the 74 sites selected.

Bioinformatics. PCR primers were trimmed from the 5′ end of individual reads using a modified version of btrim[29]. Trimmed sequences were aligned to the GRCh37 human reference genome using gapped quality-aware alignment, and base call quality recalibration implemented in Novoalign V2.07.18 with settings '-t 100 -H -g 65 -x 7 -o FullNW'. Data were realigned against known (1000 Genomes and Mills-Devine 2-hit) indels and per-sample called indels. SNPs were called using GATK 1.6-5 and settings '–min_base_quality_score 15 -stand_call_conf 30–baq CALCULATE_AS_NECESSARY -glm SNP–baqGapOpenPenalty 65–downsampling_type BY_SAMPLE–downsample_to_coverage 250' and then hard filtered using GATK settings 'QUAL<80.0 DP<20 MQ<40.0 QD<2.0 MQRankSum<-12.5 HRun>5′ (several other recommended best practice GATK settings were not appropriate for PCR amplicon data), and around indels. Small indels (up to 15-bp gaps from Novoalign) were called using GATK and settings '–min_base_quality_score 15 -stand_call_conf 30–baq CALCULATE_AS_NECESSARY -glm INDEL–baqGapOpenPenalty 65–downsampling_type BY_SAMPLE–downsample_to_coverage 250' and then hard filtered using GATK settings 'QUAL<80.0 DP<20 QD<2.0' (several other recommended best-practice GATK settings were not appropriate for PCR amplicon data). The most important of these settings were likely to be calling genotypes as missing with sequencing depth <20 high-quality bases and the minimum Phred 15 recalibrated base call quality score to define high-quality bases. Both SAMtools and VCFtools software were also used to process data. SNP genotypes (including non-reference genotypes) were called at all 58,550 bases of amplicon sequence. Samples with <57,600 SNP genotype calls (98.4%, a threshold determined by inspection of the call rate plot) were removed and scheduled for repeat processing. Clusters of very close non-reference genotypes in an individual sample were removed. Non-reference genotype sites were then identified across all samples, and VCF-level data reduced to variants at polymorphic sites (in one or more samples). A combined VCF file of all polymorphic sites and samples was then loaded into PLINK/SEQ v0.09. Multiple-step filtering based on call rate per sample and call rate per variant site was applied, with final requirements >99.95% call rate per sample and per variant site. Lower call rate samples at this stage were also scheduled for repeat processing. We removed variants if the sum of heterozygote genotype allele depths was <25% or >75%. The final filtered data was then exported to a VCF file containing all variants and samples for analysis in R. ImmunoChip data was loaded into Illumina GenomeStudio software from .idat files, and all samples called together in GenomeStudio using the cluster settings as previously described[19]. Data were merged with HapMap Phase 3 genotypes, principal component analysis performed, and the first two principal components used to validate ethnicity (Supplementary Fig. 3).

Barcode and sequencing amplicon performance. Barcode evenness was excellent, with typically 99.0% of the 1,536 barcodes producing pass-filter read numbers that were between 0.033% and 0.13% of the total pass-filter reads per lane (0.065% expected), with most of the failing barcodes tagging known water-negative control samples or (based on repeat amplification with a different barcode) due to poor DNA quality. Amplicon evenness was good, and for many genotype calls we were required to downsample data to 250 bases per site per sample (Supplementary Fig. 1). However, 10 of 511 amplicons effectively failed PCR. In a typical analysis of 100 high-quality samples, 2% of the 58,550 unique amplicon bases had a minimum mean read-depth of <20, nearly all accounted for by the 10 failing amplicons.

Variant annotation. Annotation of all variants was first performed using ANNOVAR (Feb 2013) and the GENCODE V14 data set. Coding variants were identified. Rare functional variants were identified based on stop, frameshift indel, nonsynonymous (SNV or 3*n* indel) or splice predictions. We performed an additional layer of annotation for high confidence loss of function mutations, using the methods described in ref. 30. The Variant Effect Predictor (VEP v2.5) tool from Ensembl was modified to produce custom annotation tags and additional loss of function (LOF) annotations. The additional LOF annotation was applied to variants which were annotated as STOP_GAINED, SPLICE_DONOR_VARIANT, SPLICE_ACCEPTOR_VARIANT, and FRAME_SHIFT and flagged if any filters failed. Filters included: LOF is the ancestral allele; exon is surrounded by non-canonical splice site (that is not AG/GT); LOF removes less than 5% of remaining protein; LOF is rescued by nearby start codon which results in less than 5% of protein truncated; transcript only has one coding exon; splice-site mutation within intron smaller than 15 bp; splice site is non-canonical OR other splice site within same intron is non-canonical; unable to determine exon/intron boundaries surrounding variant. A LOF variant is predicted as high confidence if there is one transcript that passes all filters, otherwise it is predicted as low confidence. We noted that LOF mutations were seen in 21 out of 25 genes, all were heterozygous genotypes, and mainly (87 out of 97) as singletons or doubletons in the 41,911 samples (Supplementary Table 3).

Statistical analysis. Most analysis was performed in R using custom code (available on request). For tests using permutations (C-alpha, UNIQ-cases and UNIQ-controls in Fig. 1), we randomly permuted in R the case–control status 10,000 times. The unconditional burden test (Fig. 1b) used a Fisher's exact text. Conditional burden tests used the glm function in R, including selected ImmunoChip common variants as covariates (selection based on a stepwise regression analysis up to 10^{-4}). For the C-alpha statistic computation (Fig. 1a), the expected proportion of rare alleles in the case–control cohorts was set to the proportion of cases and controls. Figure 1 was generated using the fact that under the null of no association $-2\log(P)$ is distributed as chi-squared with 2 degrees of freedom. PLINK/SEQ v0.09 (http://atgu.mgh.harvard.edu/plinkseq/index.shtml) was used for Ti/Tv statistics, and to confirm findings of R analyses (not shown). We used PLINK/SEQ for the genotype concordance analysis between ImmunoChip and Fluidigm-sequencing data. Discordant calls were observed at 169 of 2,985,255 (0.0056%) genotypes, occurring at 36 out of 91 polymorphic variant sites present in both data sets. We inspected Illumina ImmunoChip R theta intensity plots for the discordant genotypes, and observed 8 discordant genotypes to be likely due to ImmunoChip data mis-clustering, and 11 discordant genotypes to be due to a third or fourth observed allele in the high-throughput sequencing data. At the sites with third and fourth alleles, we note the ImmunoChip array assays can

only call two alleles, therefore is not possible to determine whether these sequence genotype calls are real or errors. R code used for analysis is available from V.P.

Estimation of average genetic effect contributed by rare variants. For each combination of locus by disease, we combined all rare functional variants (frequency $< 0.5\%$ in 1,000 Genomes/NHLBI data sets and nonsynonymous, LOF or splicing) in a burden statistic X and computed the combined frequency of X in the sample. Using a logistic regression model with the disease phenotype as outcome, we estimated the odds ratio associated with the burden variable X. This knowledge of frequency and odds ratio for the burden variable X enables the estimation of the average genetic effect (AGE, as defined in ref. 23) version of the variance explained. We then compared this variance at each combination of locus/gene with the variance explained by what we consider to be a typical common variant association (odds ratio 1.2, MAF 20%, assuming a single common variant per locus). To deal with the uncertainty in estimated odds ratio and obtain a confidence interval for this value, we randomly sampled the odds ratio from their estimated distribution for each pair of disease/locus. Averaging over the 150 combinations of 6 diseases by 25 loci, we estimate the ratio of heritability explained for all rare variants by all common variants to have a mean value of 1.6%, with a confidence interval of (1.2–2.3%). It is pointed out in ref. 23 that the AGE estimate can underestimate the true explained variance by rare variants. Nevertheless, assuming that rare variants are generally all risk or all protective at a given gene, their simulations also show that the underestimation is limited, in the range of a 25% decrease. Taking this conservative estimate of the under-estimation level, we find the upper bound of the 95% of the confidence interval to be 3.05%. Hence, our data indicate that the aggregate contribution of rare variants to the heritability ($<0.5\%$ MAF, and averaged over these loci/diseases) is unlikely to exceed approximately 3% of the heritability assigned to common variants. We acknowledge that a much larger underestimation (and therefore a much larger heritability explained for rare variants) is possible in the presence of a combination of high risk and highly protective rare variants at the same locus. Although we cannot exclude such scenario, it is unlikely to be widespread. We also assumed in our estimates that rare variants act additively at the log scale. Although this assumption is standard, we cannot exclude that a combination of rare variants results in a much stronger predictive outcome than rare variants individually, hence underestimating the heritability associated with rare variants.

24. Cooper, J. D. *et al.* Seven newly identified loci for autoimmune thyroid disease. *Hum Mol Genet* **21,** 5202–5208 (2012).
25. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491,** 119–124 (2012).
26. Sawcer, S. *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476,** 214–219 (2011).
27. Tsoi, L. C. *et al.* Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nat Genet* **44,** 1341–1348 (2012).
28. Dendrou, C. A. *et al.* Cell-specific protein phenotypes for the autoimmune locus IL2RA using a genotype-selectable human bioresource. *Nat Genet* **41,** 1011–1015 (2009).
29. Kong, Y. Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. *Genomics* **98,** 152–153 (2011).
30. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335,** 823–828 (2012).