ARTICLES

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

There is increasing evidence that genome-wide association (GWA) studies represent a powerful approach to the identification of genes involved in common human diseases. We describe a joint GWA study (using the Affymetrix GeneChip 500K Mapping Array Set) undertaken in the British population, which has examined ~2,000 individuals for each of 7 major diseases and a shared set of ~3,000 controls. Case-control comparisons identified 24 independent association signals at $P < 5 \times 10^{-7}$: 1 in bipolar disorder, 1 in coronary artery disease, 9 in Crohn's disease, 3 in rheumatoid arthritis, 7 in type 1 diabetes and 3 in type 2 diabetes. On the basis of prior findings and replication studies thus-far completed, almost all of these signals reflect genuine susceptibility effects. We observed association at many previously identified loci, and found compelling evidence that some loci confer risk for more than one of the diseases studied. Across all diseases, we identified a large number of further signals (including 58 loci with single-point *P* values between 10^{-5} and 5×10^{-7}) likely to yield additional susceptibility loci. The importance of appropriately large samples was confirmed by the modest effect sizes observed at most loci identified. This study thus represents a thorough validation of the GWA approach. It has also demonstrated that careful use of a shared control group represents a safe and effective approach to GWA analyses of multiple disease phenotypes; has generated a genome-wide genotype database for future studies of common diseases in the British population; and shown that, provided individuals with non-European ancestry are excluded, the extent of population stratification in the British population is generally modest. Our findings offer new avenues for exploring the pathophysiology of these important disorders. We anticipate that our data, results and software, which will be widely available to other investigators, will provide a powerful resource for human genetics research.

Despite extensive research efforts for more than a decade, the genetic basis of common human diseases remains largely unknown. Although there have been some notable successes[1], linkage and candidate gene association studies have often failed to deliver definitive results. Yet the identification of the variants, genes and pathways involved in particular diseases offers a potential route to new therapies, improved diagnosis and better disease prevention. For some time it has been hoped that the advent of genome-wide association (GWA) studies would provide a successful new tool for unlocking the genetic basis of many of these common causes of human morbidity and mortality[1].

Three recent advances mean that GWA studies that are powered to detect plausible effect sizes are now possible[2]. First, the International HapMap resource[3], which documents patterns of genome-wide variation and linkage disequilibrium in four population samples, greatly facilitates both the design and analysis of association studies. Second, the availability of dense genotyping chips, containing sets of hundreds of thousands of single nucleotide polymorphisms (SNPs) that provide good coverage of much of the human genome, means that for the first time GWA studies for thousands of cases and controls are technically and financially feasible. Third, appropriately large and well-characterized clinical samples have been assembled for many common diseases.

The Wellcome Trust Case Control Consortium (WTCCC) was formed with a view to exploring the utility, design and analyses of GWA studies. It brought together over 50 research groups from the UK that are active in researching the genetics of common human diseases, with expertise ranging from clinical, through genotyping, to informatics and statistical analysis. Here we describe the main experiment of the consortium: GWA studies of 2,000 cases and 3,000 shared controls for 7 complex human diseases of major public health importance—bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D). Two further experiments undertaken by the consortium will be reported elsewhere: a GWA study for tuberculosis in 1,500 cases and 1,500 controls, sampled from The Gambia; and an association study of 1,500 common controls with 1,000 cases for each of breast cancer, multiple sclerosis, ankylosing spondylitis and autoimmune thyroid disease, all typed at around 15,000 mainly non-synonymous SNPs. By simultaneously studying seven diseases with differing aetiologies, we hoped to develop insights, not only into the specific genetic contributions to each of the diseases, but also into differences in allelic architecture across the diseases. A further major aim was to address important methodological issues of relevance to all GWA studies, such as quality control, design and analysis. In addition to our main association results, we address several of these issues below, including the choice of controls for genetic studies, the extent of population structure within Great Britain, sample sizes necessary to detect genetic effects of varying sizes, and improvements in genotype-calling algorithms and analytical methods.

Samples and experimental analyses

Individuals included in the study were living within England, Scotland and Wales ('Great Britain') and the vast majority had

*Lists of participants and affiliations appear at the end of the paper.

self-identified themselves as white Europeans (153 individuals with non-Caucasian ancestry were excluded from final analysis—see below). The seven conditions selected for study are all common familial diseases of major public health importance both in the UK and globally[4], and for which suitable nationally representative sample sets were available. The control individuals came from two sources: 1,500 individuals from the 1958 British Birth Cohort (58C) and 1,500 individuals selected from blood donors recruited as part of this project (UK Blood Services (UKBS) controls). See Methods and Supplementary Table 1 for sample recruitment, phenotypes and summary details for each collection.

We adopted an experimental design with 2,000 cases for each disease and 3,000 combined controls. All 17,000 samples were genotyped with the GeneChip 500K Mapping Array Set (Affymetrix chip), which comprises 500,568 SNPs, as described in Methods. The power of this study (estimated from simulations that mimic linkage disequilibrium patterns in the HapMap Caucasian sample (CEU), see Methods) averaged across SNPs with minor allele frequencies (MAFs) above 5% is estimated to be 43% for alleles with a relative risk of 1.3, increasing to 80% for a relative risk of 1.5, for a *P*-value threshold of 5×10^{-7} (Supplementary Table 2).

We developed a new algorithm, CHIAMO, which we applied to simultaneously call the genotypes from all individuals (see Methods and Supplementary Information). Cross-platform comparison showed CHIAMO to outperform BRLMM (the standard Affymetrix algorithm) by having an error rate under 0.2% (Supplementary Table 3), and comparison of 10^8 duplicate genotypes in our study gave a discordance rate of 0.12%.

We excluded 809 samples after checks for contamination, false identity, non-Caucasian ancestry and relatedness (see Methods and Supplementary Table 4); 16,179 individuals remained in the study.

Genome-wide, 469,557 SNPs (93.8%) passed our quality control filters (described in Methods) giving an average call rate of 99.63%. Of those, 392,575 have study-wide MAFs $>1\%$ (45,106 have MAFs $<$ 0.1%; see also Supplementary Figs 1 and 2). Initial analyses of the polymorphic SNPs suggest that patterns of linkage disequilibrium in our samples are very similar to those in HapMap (Supplementary Fig. 3). Therefore, we expect genome coverage with the Affymetrix 500K set in this study to be similar to that estimated for the HapMap CEU panel[2].

All SNPs passing quality control filters were used in the association analyses, although power is very low for SNPs with low MAFs (unless they have unusually large effects). On visual inspection of the cluster plots of SNPs showing apparently strong association, we removed a further 638 SNPs with poor clustering.

Control groups

Our main purpose in using two control groups was to assess possible bias in ascertaining control samples. In addition, noting that DNA sample processing differed between these groups, comparison of control groups also provides a check for effects of differential genotyping errors as a result of differences in DNA collection and preparation. Figure 1a shows the results of 1-d.f. Mantel-extension tests[5] for differences in allele frequencies of SNPs between subjects from the 58BC and UKBS collections, stratified by 12 broad regions of Great Britain (see Supplementary Table 5 and Supplementary Fig. 4 for region definitions). The associated quantile-quantile plot (see Methods for background) in Fig. 1b shows good agreement with the null distribution (similar results are obtained for tests that do not stratify by geography, data not shown). The fact that we see few significant differences between these two control groups despite the fact that they differ in population groups sampled, DNA processing, and age, indicates that there would be little bias due to use of either sample as a control group for any of the case series, and justifies our combining of the two control groups to form a single group of 3,000 subjects for our main analyses.

One consequence of using a shared control group (for which detailed phenotyping for all traits of interest is not available) relates to the potential for misclassification bias: a proportion of the controls is likely to have the disease of interest (and therefore might meet the criteria for inclusion as a case) and some others will develop it in the future. However, the effect this has on power is modest unless the extent of misclassification bias is substantial; for example, if 5% of controls would meet the definition of cases at the same age, the loss of power is approximately the same as that due to a reduction of the sample size by 10%[6]. Even for the higher prevalence conditions examined by the WTCCC (such as HT, CAD and T2D), the precise ascertainment schemes used here (which enriched for more extreme phenotypes and/or strong family history) will have limited the proportions of controls meeting case criteria to low levels (for example, to $<5\%$). Although a study design which used 'hypercontrols' (that is, selection of control individuals from the lower extremity of the relevant trait distribution) would generally be the most powerful approach in a study focusing on one disease, the merits of such an approach need to be weighed against the additional costs associated with the need to phenotype and genotype each control sample.

Geographical variation and population structure

An additional cause of false positive findings is hidden population structure. Case and control samples may differ in the distribution of their ancestry, either owing to control sampling effects, as discussed above, or to confounding when different ancestries carry higher disease risk and are, as a result, over-represented in cases. Even after exclusion of individuals with evidence of recent non-European ancestry, the British population is heterogeneous, having been shaped by several waves of immigration from southern and northern Europe. Whether the differences between these incoming populations are sufficiently large to distort the findings of population-based case-control studies is an open question.

We first examined our samples for non-European ancestry, using multidimensional scaling after 'seeding' our data with those from the three HapMap analysis panels (see Supplementary Fig. 5 and

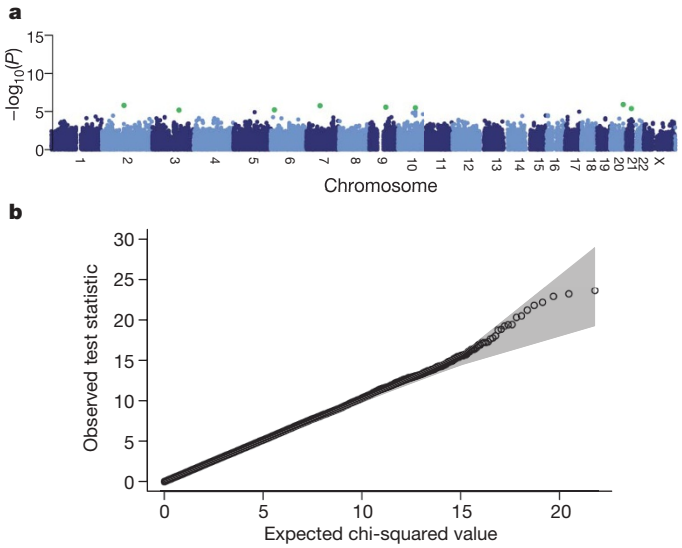


Figure 1 | Genome-wide scan for allele frequency differences between controls. **a**, *P* values from the trend test for differences between SNP allele frequencies in the two control groups, stratified by geographical region. SNPs have been excluded on the basis of failure in a test for Hardy–Weinberg equilibrium in either control group considered separately, a low call rate, or if minor allele frequency is less than 1%, but not on the basis of a difference between control groups. Green dots indicate SNPs with a *P* value $<1 \times 10^{-5}$. **b**, Quantile-quantile plots of these test statistics. In this and subsequent quantile-quantile plots, the shaded region is the 95% concentration band (see Methods).

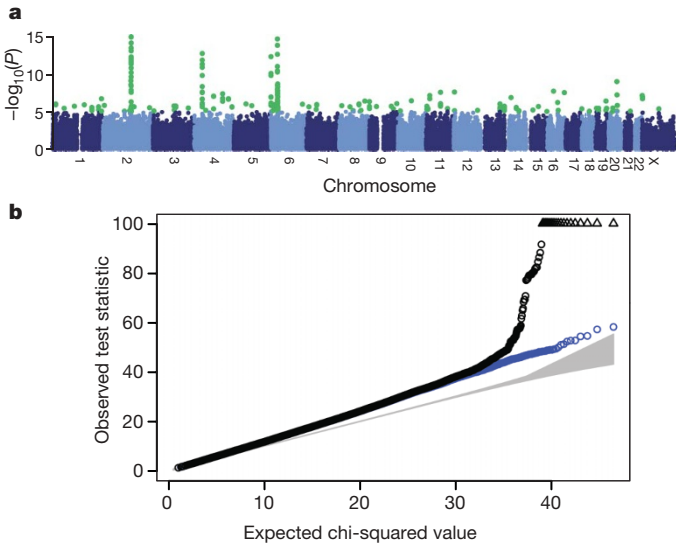


Figure 2 | Genome-wide picture of geographic variation. a, *P* values for the 11-d.f. test for difference in SNP allele frequencies between geographical regions, within the 9 collections. SNPs have been excluded using the project quality control filters described in Methods. Green dots indicate SNPs with a *P* value $<1 \times 10^{-5}$. **b**, Quantile-quantile plots of these test statistics. SNPs at which the test statistic exceeds 100 are represented by triangles at the top of the plot, and the shaded region is the 95% concentration band (see Methods). Also shown in blue is the quantile-quantile plot resulting from removal of all SNPs in the 13 most differentiated regions (Table 1).

Methods), and excluded 153 individuals on this basis. We next looked for evidence of population heterogeneity by studying allele frequency differences between the 12 broad geographical regions (defined in Supplementary Fig. 4). The results for these 11-d.f. tests and associated quantile-quantile plots are shown in Fig. 2. Widespread small differences in allele frequencies are evident as an increased slope of the line (Fig. 2b); in addition, a few loci show much larger differences (Fig. 2a and Supplementary Fig. 6).

Thirteen genomic regions showing strong geographical variation are listed in Table 1, and Supplementary Fig. 7 shows the way in which their allele frequencies vary geographically. The predominant pattern is variation along a NW/SE axis. The most likely cause for these marked geographical differences is natural selection, most plausibly in populations ancestral to those now in the UK. Variation due to selection has previously been implicated at *LCT* (lactase) and major histocompatibility complex (MHC)[7–9], and within-UK differentiation at 4p14 has been found independently[10], but others seem to be new findings. All but three of the regions contain known genes. Aside from evolutionary interest, genes showing evidence of natural selection are particularly interesting for the biology of traits such as infectious diseases; possible targets for selection include *NADSYN1* (NAD synthetase 1) at 11q13, which could have a role in prevention of pellagra, as well as *TLR1* (toll-like receptor 1) at 4p14, for which a role in the biology of tuberculosis and leprosy has been suggested[10].

There may be important population structure that is not well captured by current geographical region of residence. Present implementations of strongly model-based approaches such as STRUCTURE[11,12] are impracticable for data sets of this size, and we reverted to the classical method of principal components[13,14], using a subset of 197,175 SNPs chosen to reduce inter-locus linkage disequilibrium. Nevertheless, four of the first six principal components clearly picked up effects attributable to local linkage disequilibrium rather than genome-wide structure. The remaining two components show the same predominant geographical trend from NW to SE but, perhaps unsurprisingly, London is set somewhat apart (Supplementary Fig. 8).

The overall effect of population structure on our association results seems to be small, once recent migrants from outside Europe are excluded. Estimates of over-dispersion of the association trend test statistics (usually denoted λ; ref. 15) ranged from 1.03 and 1.05 for RA and T1D, respectively, to 1.08–1.11 for the remaining diseases. Some of this over-dispersion could be due to factors other than structure, and this possibility is supported by the fact that inclusion of the two ancestry informative principal components as covariates in the association tests reduced the over-dispersion estimates only slightly (Supplementary Table 6), as did stratification by geographical region. This impression is confirmed on noting that *P* values with and without correction for structure are similar (Supplementary Fig. 9). We conclude that, for most of the genome, population structure has at most a small confounding effect in our study, and as a consequence the analyses reported below do not correct for structure. In principle, apparent associations in the few genomic regions identified in Table 1 as showing strong geographical differentiation should be interpreted with caution, but none arose in our analyses.

Disease association results

We assessed evidence for association in several ways (see Methods for details), drawing on both classical and bayesian statistical approaches. For polymorphic SNPs on the Affymetrix chip, we performed trend tests (1 degree of freedom[16]) and general genotype tests (2 degrees of freedom[16], referred to as genotypic) between each case collection and the pooled controls, and calculated analogous Bayes factors. There are examples from animal models where genetic effects act differently in males and females[17], and to assess this in our data we applied a

Table 1 | Highly differentiated SNPs

Chromosome	Genes	Region (Mb)	SNP	Position	*P* value
2q21	*LCT*	135.16–136.82	rs1042712	136,379,576	5.54×10^{-13}
4p14	*TLR1, TLR6, TLR10*	38.51–38.74	rs7696175	386,43,552	1.51×10^{-12}
4q28		137.97–138.01	rs1460133	137,999,953	4.43×10^{-08}
6p25	*IRF4*	0.32–0.42	rs9378805	362,727	5.39×10^{-13}
6p21	*HLA*	31.10–31.55	rs3873375	31,359,339	1.07×10^{-11}
9p24	*DMRT1*	0.86–0.88	rs11790408	866,418	4.96×10^{-07}
11p15	*NAV2*	19.55–19.70	rs12295525	19,661,808	7.44×10^{-08}
11q13	*NADSYN1, DHCR7*	70.78–70.93	rs12797951	70,820,914	3.01×10^{-08}
12p13	*DYRK4,AKAP3,NDUFA9, RAD51AP1,GALNT8*	4.37–4.82	rs10774241	45,537,27	2.73×10^{-08}
14q12	*HECTD1,AP4S1,STRN3*	30.41–31.03	rs17449560	30,598,823	1.46×10^{-07}
19q13	*GIPR,SNRPD2,QPCTL, SIX5,DMPK,DMWD, RSHL1,SYMPK,FOXA3*	50.84–51.09	rs3760843	50,980,546	4.19×10^{-07}
20q12		38.30–38.77	rs2143877	38,526,309	1.12×10^{-09}
Xp22		2.06–2.08	rs6644913	2,061,160	1.23×10^{-07}

Properties of SNPs that show large allele frequency differences between samples of individuals from 12 regions across Great Britain. Regions showing differentiated SNPs are given with details of the SNP with the smallest *P* value in each region for differentiation on the 11-d.f. test of differences in SNP allele frequencies between geographical regions, within the 9 collections. Cluster plots for these SNPs have been examined visually. Signal plots appear in Supplementary Information. Positions are in NCBI build-35 coordinates.

Box 1 | Significance levels in genome-wide studies

There has been much debate concerning interpretation of significance levels in genome-wide association studies and whether, and how, these should be corrected for multiple testing. Classical multiple testing theory in statistics is concerned with the problem of 'multiple tests' of a single 'global' null hypothesis. This, we would argue, is a problem far removed from that which faces us in genome-wide association studies, where we face the problem of testing 'multiple hypotheses' (for a particular disease, one hypothesis for each SNP, or region of correlated SNPs, in the genome) and we thus do not subscribe to the view that one should correct significance levels for the number of tests performed to obtain 'genome-wide significance levels'. Nonetheless, our aim is to keep the false positive rate within acceptable bounds and this still leads to the view that very low *P* values are needed for strong evidence of association. But the factor determining the threshold is not the number of tests performed, but the a priori probability that there is likely to be a true association at any specified location in the genome. Of course, we cannot know this prior probability from objective evidence, but we can perhaps estimate an order of magnitude.

There are two linked questions. The first concerns the choice of an appropriate 'threshold' for reporting possible associations as likely to be genuine. Here the mathematics is quite straightforward if we make the simplifying assumption that we have the same power to detect all true associations. Then we have[18]

Posterior odds for true association = Prior odds × Power/Significance threshold

That is, for a given significance threshold, the probability of a true association depends on the prior odds and, crucially, the power. A plausible estimate for the prior odds of true association at any specified locus might be of the order of 100,000:1 against, for example, on the basis of 1,000,000 'independent' regions of the genome and an expectation of 10 detectable genes involved in the condition. (Other plausible estimates might vary from this by an order of magnitude or so in either direction.) Then, assuming a power of 0.5 and a significance threshold of 5×10^{-7}, the posterior odds in favour of a 'hit' being a true association would be 10:1. However, if we relax this significance threshold by a factor of ten, or alternatively if the power were lower by a factor of 10, the posterior odds that a 'hit' is a true association would also be reduced by a factor of ten. This simple mathematical analysis is little affected by allowing for the fact that true associations come in various sizes with varying power to detect them; the above formula is simply modified by interpreting 'power' as the mean power.

The above discussion concerns 'average' properties of 'hits' achieving given significance levels. After the association data are available, a related but different question is whether a particular positive finding is likely to be a true one. For that calculation, the prior odds must be multiplied by the Bayes factor, the ratio of the probability of the observed data under the assumption that there is a true association to its probability under the null hypothesis. As in power calculations, the calculation of Bayes factors requires assumptions about effect sizes (see Methods for details).

A key point from both perspectives is that interpreting the strength of evidence in an association study depends on the likely number of true associations, and the power to detect them which, in turn, depends on effect sizes and sample size. In a less-well-powered study it would be necessary to adopt more stringent thresholds to control the false-positive rate. Thus, when comparing two studies for a particular disease, with a hit with the same MAF and *P* value for association, the likelihood that this is a true positive will in general be greater for the study that is better powered, typically the larger study. In practice, smaller studies often employ less stringent *P*-value thresholds, which is precisely the opposite of what should occur.

sex-differentiated test which is sensitive to associations of a different magnitude and/or direction in the two sexes.

Our study also allows us to look for loci which may have an effect in more than one disease. To assess this, we compared our common controls with all cases in each of three natural groupings of diseases: CAD+HT+T2D (metabolic and cardiovascular phenotypes with potential aetiological overlap, for example, involving defects in insulin action); RA+T1D (already known to share common loci); and CD+RA+T1D (all autoimmune diseases).

To help to capture putative disease loci not on the Affymetrix chip we used a new multilocus method in which a population genetics model is applied to our genotype data and the HapMap reference samples to simulate, or impute, genotype data at 2,193,483 HapMap SNPs not on the Affymetrix chip. These imputed, or *in silico*, genotypes are then tested for association in the same ways as SNPs genotyped in the project.

Before detailing the principal results for each disease, we first summarize our main observations. Table 2 details the findings from the WTCCC scan for the 15 variants for which there was strong prior evidence of association with one or more of the diseases studied, based on extensive replication studies. All but two of these show associations in our study, with the magnitude of the evidence generally consistent with their effect sizes as estimated from prior studies. One of the signals for which we failed to obtain evidence of replication (*APOE* in CAD) is poorly tagged by the Affymetrix 500K chip. The other (*INS* in T1D) is represented by a single SNP that marginally failed our study-wide quality control filters (overall missingness 5.2%) but which was nonetheless strongly associated with T1D when examined. Quantile-quantile plots for the trend test for each of the seven diseases show only very minor deviations from the null distribution, except in the extreme tails which correspond to associations reported below (Fig. 3). The quantile-quantile plots and the results at positive controls (Table 2) give confidence in the quality of our data and the robustness of our analyses.

Our genome-wide results for the trend test are illustrated in Fig. 4. The single-disease trend and genotypic tests for SNPs on the chip identified 21 signals across the 7 diseases that exceeded a threshold of 5×10^{-7} (Table 3). For each of these SNPs (except those within the MHC), cluster plots are shown in Supplementary Fig. 10 and 'signal plots' in Fig. 5. These signal plots estimate the likely demarcation of the hit region and show the signal at genotyped and imputed SNPs together with local genomic context. Four further strong (with $P < 5 \times 10^{-7}$) associations were revealed by the other primary analyses described (Table 3). One locus (in RA) was revealed by the sex-differentiated analysis, two through multilocus approaches (both for T1D) and one through an analysis which combined cases from more than one autoimmune disease (signal plots in Supplementary Figs 11, 12 and 13, respectively).

All of these signals were subjected to visual inspection of cluster plots, and in all cases (with one exception noted below) nearby correlated SNPs also showed a strong signal (see signal plots). Thus, genotyping artefacts are unlikely to be responsible for these associations. Indeed, at the time of writing, 12 of these 25 strong signals represent replications of previously reported findings (only those with extensive prior replication are reported in Table 2). Of the remainder, follow-up studies (reported elsewhere) have confirmed all but one of the loci (ten in total) for which replication has been attempted[10,19–24]. The other replication study gave equivocal results. Of the 18 loci implicated in autoimmune diseases, 5 show associations ($P < 0.001$) to more than 1 condition, leading to a number of further potential new associations, at least one of which has also been replicated[10].

It is likely that further susceptibility genes will be identified through follow-up of other signals for which the evidence from our scan is less conclusive (see below for some specific examples). For example, there are 58 further signals with single-point *P* values between 10^{-5} and 5×10^{-7} for which inspection of cluster plots verifies CHIAMO calls (Table 4). As described below, analyses which make use of selected case samples to expand the reference group should also provide a useful route to the prioritization of such putative signals for further analysis. For convenience, the strongest association results are presented separately for each disease in Supplementary Table 7.

Several general points are relevant to interpretation of these disease-association data. First, replication studies are required to confirm associations from GWAs. For the reasons given in the box, we regard very low *P* values (say $P < 5 \times 10^{-7}$) in our comparatively large sample size as strong evidence for association, and indeed all

Table 2 | Evidence for signal of association at previously robustly replicated loci

Collection	Gene	Chromosome	Reported SNP	WTCCC SNP	HapMap r^2	Trend *P* value	Genotypic *P* value
CAD	*APOE*	19q13	*	rs4420638	-	1.7×10^{-01}	1.7×10^{-01}
CD	*NOD2*	16q12	rs2066844	rs17221417	0.23	9.4×10^{-12}	4.0×10^{-11}
CD	*IL23R*	1p31	rs11209026	rs11805303	0.01	6.5×10^{-13}	5.9×10^{-12}
RA	*HLA-DRB1*	6p21	*	rs615672	-	2.6×10^{-27}	7.5×10^{-27}
RA	*PTPN22*	1p13	rs2476601	rs6679677	0.75	4.9×10^{-26}	5.6×10^{-25}
T1D	*HLA-DRB1*	6p21	*	rs9270986	-	4.0×10^{-116}	2.3×10^{-122}
T1D	*INS*	11p15	rs689	†	-	-	-
T1D	*CTLA4*	2q33	rs3087243	rs3087243	1	2.5×10^{-05}	1.8×10^{-05}
T1D	*PTPN22*	1p13	rs2476601	rs6679677	0.75	1.2×10^{-26}	5.4×10^{-26}
T1D	*IL2RA*	10p15	rs706778	rs2104286	0.25	8.0×10^{-06}	4.3×10^{-05}
T1D	*IFIH1*	2q24	rs1990760	rs3788964	0.26	1.9×10^{-03}	7.6×10^{-03}
T2D	*PPARG*	3p25	rs1801282	rs1801282	1	1.3×10^{-03}	5.4×10^{-03}
T2D	*KCNJ11*	11p15	rs5219	rs5215	0.9	1.3×10^{-03}	5.6×10^{-03}
T2D	*TCF7L2*	10q25	rs7903146	rs4506565	0.92	5.7×10^{-13}	5.1×10^{-12}

Where information on the strength of association at a particular SNP had been previously published and replicated we tabulated the *P* value of both the trend and genotype test at the same SNP (if in our study), or the best tag SNP (defined to be the SNP with highest r^2 with the reported SNP, calculated in the CEU sample of the HapMap project). Positions are in NCBI build-35 coordinates. *Previous reports relate to haplotypes rather than single SNPs. †Not well tagged by SNPs that pass the quality control, see main text.

or most of the loci we find at this level are either already known or have now been confirmed by subsequent replication. Such replication studies are also the substrate for efforts to determine the range of associated phenotypes and to identify and characterize pathologically relevant variation.

Second, failure to detect a prominent association signal in the present study cannot provide conclusive exclusion of any given gene. This is the consequence of several factors including: less-than-complete coverage of common variation genome-wide on the Affymetrix chip; poor coverage (by design) of rare variants, including many structural variants (thereby reducing power to detect rare, penetrant, alleles)[25]; difficulties with defining the full genomic extent of the gene of interest; and, despite the sample size, relatively low power to detect, at levels of significance appropriate for genome-wide analysis, variants with modest effect sizes (odds ratio (OR) < 1.2).

Third, whereas the association signals detected can help to define regions of interest, they cannot provide unambiguous identification of the causal genes. Nevertheless, assessments on the basis of positional candidacy carry considerable weight, and, as we show, these already allow us, for selected diseases, to highlight pathways and mechanisms of particular interest. Naturally, extensive resequencing and fine-mapping work, followed by functional studies will be required before such inferences can be translated into robust statements about the molecular and physiological mechanisms involved.

We turn now to a discussion of the main findings for each disease, focusing here only on the most significant and interesting results

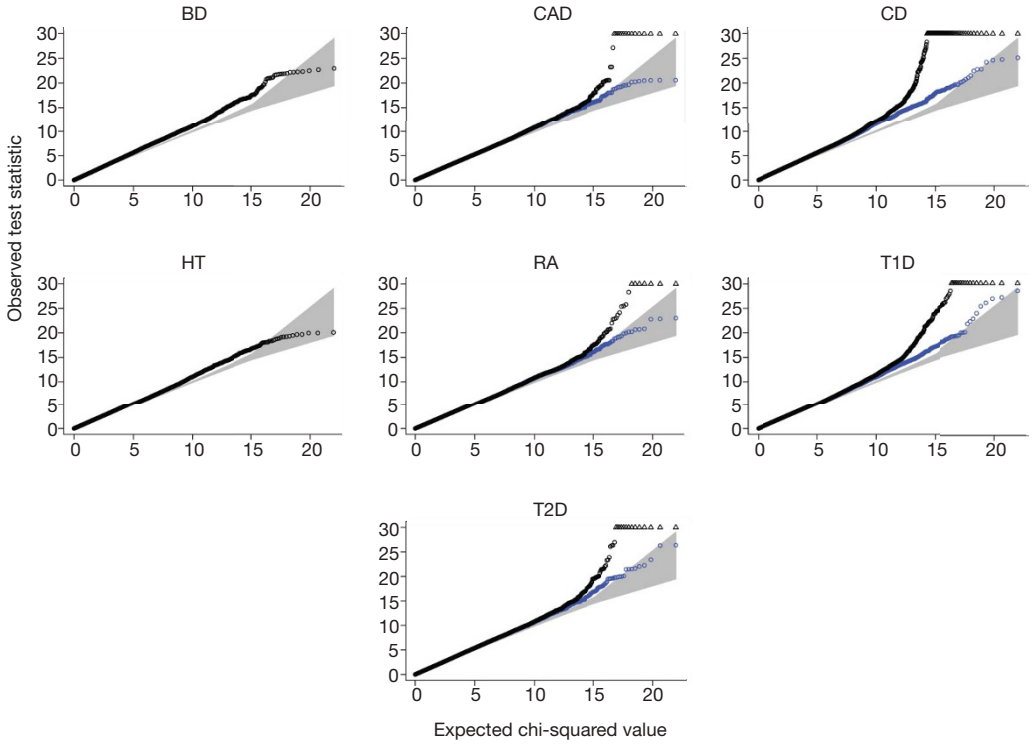


Figure 3 | Quantile-quantile plots for seven genome-wide scans. For each of the seven disease collections, a quantile-quantile plot of the results of the trend test is shown in black for all SNPs that pass the standard project filters, have a minor allele frequency >1% and missing data rate <1%. SNPs that were visually inspected and revealed genotype calling problems were excluded. These filters were chosen to minimize the influence of genotype-calling artefacts. Each quantile-quantile plot shown in black involves around 360,000 SNPs. SNPs at which the test statistic exceeds 30 are represented by triangles. Additional quantile-quantile plots, which also exclude all SNPs located in the regions of association listed in Table 3, are superimposed in blue (for BD, the exclusion of these SNPs has no visible effect on the plot, and for HT there are no such SNPs). The blue quantile-quantile plots show that departures in the extreme tail of the distribution of test statistics are due to regions with a strong signal for association.

from the analyses described above, and consideration of an expanded reference group, described below.

Bipolar disorder (BD). Bipolar disorder (BD; manic depressive illness[26]) refers to an episodic recurrent pathological disturbance in mood (affect) ranging from extreme elation or mania to severe depression and usually accompanied by disturbances in thinking and behaviour: psychotic features (delusions and hallucinations) often occur. Pathogenesis is poorly understood but there is robust evidence for a substantial genetic contribution to risk[27,28]. The estimated sibling recurrence risk (λ_s) is 7–10 and heritability 80–90%[27,28]. The definition of BD phenotype is based solely on clinical features because, as yet, psychiatry lacks validating diagnostic tests such as those available for many physical illnesses. Indeed, a major goal of molecular genetics approaches to psychiatric illness is an improvement in diagnostic classification that will follow identification of the biological systems that underpin the clinical syndromes. The phenotype definition that we have used includes individuals that have suffered one or more episodes of pathologically elevated mood (see Methods), a criterion that captures the clinical spectrum of bipolar mood variation that shows familial aggregation[29].

Several genomic regions have been implicated in linkage studies[30] and, recently, replicated evidence implicating specific genes has been reported. Increasing evidence suggests an overlap in genetic susceptibility with schizophrenia, a psychotic disorder with many similarities to BD. In particular association findings have been reported with both disorders at *DAOA* (D-amino acid oxidase activator), *DISC1* (disrupted in schizophrenia 1), *NRG1* (neuregulin1) and *DTNBP1* (dystrobrevin binding protein 1)[31].

The strongest signal in BD was with rs420259 at chromosome 16p12 (genotypic test $P = 6.3 \times 10^{-8}$; Table 3) and the best-fitting genetic model was recessive (Supplementary Table 8). Although recognizing that this signal was not additionally supported by the expanded reference group analysis (see below and Supplementary Table 9) and that independent replication is essential, we note that several genes at this locus could have pathological relevance to BD, (Fig. 5). These include *PALB2* (partner and localizer of *BRCA2*), which is involved in stability of key nuclear structures including chromatin and the nuclear matrix; *NDUFAB1* (NADH dehydrogenase (ubiquinone) 1, alpha/beta subcomplex, 1), which encodes a subunit of complex I of the mitochondrial respiratory chain; and *DCTN5* (dynactin 5), which encodes a protein involved in intracellular transport that is known to interact with the gene 'disrupted in schizophrenia 1' (*DISC1*)[32], the latter having been implicated in susceptibility to bipolar disorder as well as schizophrenia[33].

Of the four regions showing association at $P < 5 \times 10^{-7}$ in the expanded reference group analysis (Supplementary Table 9), it is of interest that the closest gene to the signal at rs1526805 ($P = 2.2 \times 10^{-7}$) is *KCNC2* which encodes the Shaw-related voltage-gated potassium channel. Ion channelopathies are well-recognized as causes of episodic central nervous system disease, including seizures, ataxias

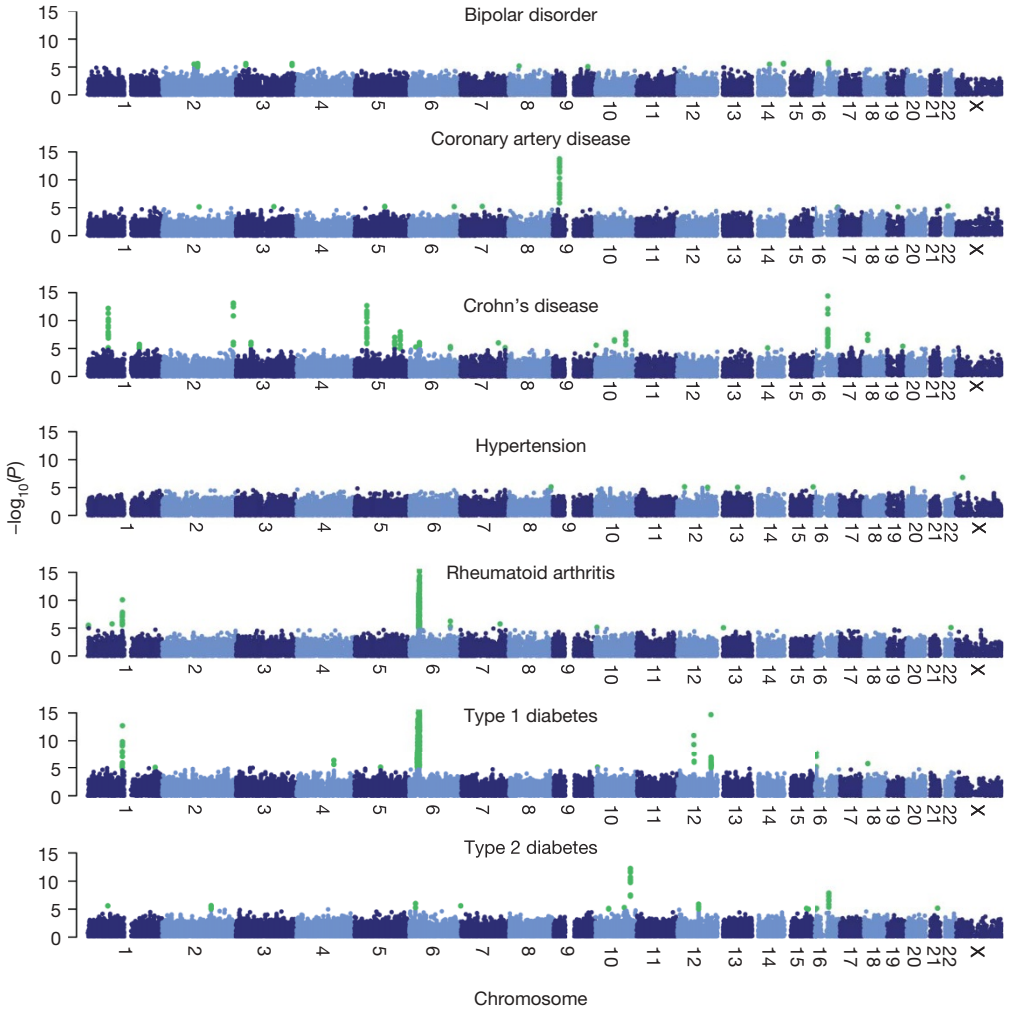


Figure 4 | Genome-wide scan for seven diseases. For each of seven diseases $-\log_{10}$ of the trend test *P* value for quality-control-positive SNPs, excluding those in each disease that were excluded for having poor clustering after visual inspection, are plotted against position on each chromosome. Chromosomes are shown in alternating colours for clarity, with *P* values $< 1 \times 10^{-5}$ highlighted in green. All panels are truncated at $-\log_{10}(P \text{ value}) = 15$, although some markers (for example, in the MHC in T1D and RA) exceed this significance threshold.

Table 3 | Regions of the genome showing the strongest association signals

Collection	Chromosome	Region (Mb)	SNP	Trend *P* value	Genotypic *P* value	$\log_{10}$(BF), additive	$\log_{10}$(BF), general	Risk allele	Minor allele	Heterozygote odds ratio	Homozygote odds ratio	Control MAF	Case MAF
Standard analysis													
BD	16p12	23.3–23.62	rs420259	2.19×10^{-04}	6.29×10^{-08}	1.96	4.79	A	G	2.08 (1.60–2.71)	2.07 (1.6–2.69)	0.282	0.248
CAD	9p21	21.93–22.12	rs1333049	1.79×10^{-14}	1.16×10^{-13}	11.66	11.19	C	C	1.47 (1.27–1.70)	1.9 (1.61–2.24)	0.474	0.554
CD	1p31	67.3–67.48	rs11805303	6.45×10^{-13}	5.85×10^{-12}	10.07	9.41	T	T	1.39 (1.22–1.58)	1.86 (1.54–2.24)	0.317	0.391
CD	2q37	233.92–234	rs10210302	7.10×10^{-14}	5.26×10^{-14}	11.11	11.28	T	C	1.19 (1.01–1.41)	1.85 (1.56–2.21)	0.481	0.402
CD	3p21	49.3–49.87	rs9858542	7.71×10^{-07}	3.58×10^{-08}	4.24	5.22	A	A	1.09 (0.96–1.24)	1.84 (1.49–2.26)	0.282	0.331
CD	5p13	40.32–40.66	rs17234657	2.13×10^{-13}	1.99×10^{-12}	10.41	9.89	G	G	1.54 (1.34–1.76)	2.32 (1.59–3.39)	0.125	0.181
CD	5q33	150.15–150.31	rs1000113	5.10×10^{-08}	3.15×10^{-07}	5.36	5.01	T	T	1.54 (1.31–1.82)	1.92 (0.92–4.00)	0.067	0.098
CD	10q21	64.06–64.31	rs10761659	2.68×10^{-07}	1.75×10^{-06}	4.69	4.13	G	A	1.23 (1.05–1.45)	1.55 (1.3–1.84)	0.461	0.406
CD	10q24	101.26–101.32	rs10883365	1.41×10^{-08}	5.82×10^{-08}	5.91	5.48	G	G	1.2 (1.03–1.39)	1.62 (1.37–1.92)	0.477	0.537
CD	16q12	49.02–49.4	rs17221417	9.36×10^{-12}	3.98×10^{-11}	8.93	8.47	G	G	1.29 (1.13–1.46)	1.92 (1.58–2.34)	0.287	0.356
CD	18p11	12.76–12.91	rs2542151	4.56×10^{-08}	2.03×10^{-07}	5.42	5.00	G	G	1.3 (1.14–1.48)	2.01 (1.46–2.76)	0.163	0.208
RA	1p13	113.54–114.16	rs6679677	4.90×10^{-26}	5.55×10^{-25}	22.36	21.99	A	A	1.98 (1.72–2.27)	3.32 (1.93–5.69)	0.096	0.168
RA	6	MHC	rs6457617*	3.44×10^{-76}	5.18×10^{-75}	74.84	73.18	T	T	2.36 (1.97–2.84)	5.21 (4.31–6.30)	0.489	0.685
T1D	1p13	113.54–114.16	rs6679677	1.17×10^{-26}	5.43×10^{-26}	23.07	22.83	A	A	1.82 (1.59–2.09)	5.19 (3.15–8.55)	0.096	0.169
T1D	6	MHC	rs9272346*	2.42×10^{-134}	5.47×10^{-134}	141.9	142.2	A	G	5.49 (4.83–6.24)	18.52 (27.03–12.69)	0.387	0.150
T1D	12q13	54.64–55.09	rs11171739	1.14×10^{-11}	9.71×10^{-11}	8.89	8.24	C	C	1.34 (1.17–1.54)	1.75 (1.48–2.06)	0.423	0.493
T1D	12q24	109.82–111.49	rs17696736	2.17×10^{-15}	1.51×10^{-14}	12.53	11.88	G	G	1.34 (1.16–1.53)	1.94 (1.65–2.29)	0.424	0.506
T1D	16p13	10.93–11.37	rs12708716	9.24×10^{-08}	4.92×10^{-07}	5.15	4.70	A	G	1.19 (0.97–1.45)	1.55 (1.27–1.89)	0.350	0.297
T2D	6p22	20.63–20.84	rs9465871	1.02×10^{-06}	3.34×10^{-07}	4.15	3.98	C	C	1.18 (1.04–1.34)	2.17 (1.6–2.95)	0.178	0.218
T2D	10q25	114.71–114.81	rs4506565	5.68×10^{-13}	5.05×10^{-12}	10.14	9.43	T	T	1.36 (1.2–1.54)	1.88 (1.56–2.27)	0.324	0.395
T2D	16q12	52.36–52.41	rs9939609	5.24×10^{-08}	1.91×10^{-07}	5.35	5.05	A	A	1.34 (1.17–1.52)	1.55 (1.3–1.84)	0.398	0.453
Multi-locus analysis													
T1D	4q27	123.26–123.92	rs6534347	4.48×10^{-07}	1.83×10^{-06}	5.15	4.69	A	A	1.30 (1.10–1.55)	1.49 (1.25–1.78)	0.351	0.402
T1D	12p13	9.71–9.86	rs3764021	7.19×10^{-05}	5.08×10^{-08}	2.12	4.55	C	T	1.57 (1.38–1.79)	1.48 (1.25–1.75)	0.467	0.426
Sex differentiated analysis													
RA	7q32	130.80–130.84	rs11761231	3.91×10^{-07}	1.37×10^{-06}	-	-	G	A	1.44 (1.19–1.75)	1.64 (1.35–1.99)	0.375	0.327
Combined cases													
RA+T1D	10p15	6.07–6.17	rs2104286	5.92×10^{-08}	2.52×10^{-07}	5.26	4.45	T	C	1.35 (1.11–1.65)	1.62 (1.34–1.97)	0.286	0.245

Regions with at least one SNP with a *P* value of less than 5×10^{-7} for our primary analyses. The $\log_{10}$ value of the Bayes factor (BF) for the bayesian analysis corresponding to the trend and genotypic tests is also given. Region marks the boundaries of signal defined by recombination and return of test statistics to background levels. The minor allele is defined in the controls and its frequency in that group as well as the case sample is reported. MAF, minor allele frequency. Cluster plots for each SNP have been inspected visually, and are shown in Supplementary Fig. 10. Positions are in NCBI build-35 coordinates *Multiple SNPs in the MHC region are significant, we report the most extreme.

and paralyses[34]. It is possible that this may extend to episodic disturbances of mood and behaviour.

Amongst the other higher ranked signals in the BD data set (Supplementary Table 7), there is support for the previously suggested importance of GABA neurotransmission (rs7680321 ($P = 6.2 \times 10^{-5}$) in *GABRB1* encoding a ligand-gated ion channel (GABA A receptor, beta 1))[35], glutamate neurotransmission (rs1485171 ($P = 9.7 \times 10^{-5}$) in *GRM7* (glutamate receptor, metabotropic 7))[35] and synaptic function (rs11089599 ($P = 7.2 \times 10^{-5}$) in *SYN3* (synapsin III)[36]).

We note that a broad range of genetic and non-genetic data point to the importance of analyses that use alternative approaches to phenotype definition, including symptom dimensions[31]. Although beyond the scope of the current paper, such analyses will be required to maximize the potential of the current BD data set.

Coronary artery disease (CAD). Coronary artery disease (coronary atherosclerosis) is a chronic degenerative condition in which lipid and fibrous matrix is deposited in the walls of the coronary arteries to form atheromatous plaques[37]. It may be clinically silent or present with angina pectoris or acute myocardial infarction. Pathogenesis is complex, with endothelial dysfunction, oxidative stress and inflammation contributing to development and instability of the atherosclerotic plaque[37].

In addition to lifestyle and environmental factors, genes are important in the aetiology of CAD[38]. For early myocardial infarction, estimates of λ_s range from ~2 to ~7 (ref. 39). Genetic variation is thought likely to influence risk of CAD both directly and through effects on known CAD risk factors including hypertension, diabetes and hypercholesterolaemia. Genome-wide linkage studies have mapped several loci that may affect susceptibility to CAD/myocardial infarction[40] although for only two of these has the likely gene been identified (*ALOX5AP* (arachidonate 5-lipoxygenase-activating protein) and *LTA4H* (leukotriene A4 hydrolase))[41,42]. Association studies have identified several plausible genetic variants affecting lipids, thrombosis, inflammation or vascular biology but for most the evidence is not yet conclusive[40]. We did not find evidence for strong association at any of these genes within our study (Table 2 and Supplementary Table 10).

The most notable new finding for CAD is the powerful association on chromosome 9p21.3 (Table 3; Fig. 5). Although the strongest signal is seen at rs1333049 ($P = 1.8 \times 10^{-14}$), associations are seen for SNPs across > 100 kilobases. This region has not been highlighted in previous studies of CAD or myocardial infarction[40,43]. The region of interest contains the coding sequences of genes for two cyclin dependent kinase inhibitors, *CDKN2A* (encoding $p16^{INK4a}$) and *CDKN2B* ($p15^{INK4b}$), although the most closely associated SNP is some distance removed. Both genes have multiple isoforms, have an important role in the regulation of the cell cycle and are widely expressed[44], with *CDKN2B* known to be expressed in the macrophages but not the smooth muscle cells of fibrofatty lesions[45,46]. It is of interest that expression of *CDKN2B* is induced by transforming growth factor beta (TGF-β) and that the TGF-β signalling system is implicated in the pathogenesis of human atherosclerosis[45,46]. Besides *CDKN2A* and *CDKN2B*, the only other known gene nearby is *MTAP* which encodes methylthioadenosine phosphorylase, an enzyme that contributes to polyamine metabolism and is important for the salvage of both adenine and methionine. *MTAP* is ubiquitously expressed, including in the cardiovascular system[47]. Further work is required to determine whether the CAD association at this locus is mediated through *CDKN2A/B, MTAP* or some other mechanism. The same region also shows replicated evidence of association to T2D in the WTCCC and other data sets[19,21,22], though different SNPs seem to be involved.

None of the loci showing more modest associations with CAD (Table 4) includes genes hitherto strongly implicated in the pathogenesis of CAD. A potentially interesting association is at rs6922269 ($P = 6.3 \times 10^{-6}$), an intronic SNP in *MTHFD1L*, which encodes

methylenetetrahydrofolate dehydrogenase ($NADP^+$-dependent) 1-like, the mitochondrial isozyme of C1-tetrahydrofolate (THF) synthase[48,49]. C_1-THF synthases interconvert the one carbon units carried by the biologically active form of folic acid, C1-tetrahydrofolate. These are used in a variety of cellular processes including purine and methionine synthesis[48]. Another enzyme in the same pathway, methylene THF reductase (encoded by *MTHFR*) is subject to a common mutation which influences plasma homocysteine level[50] and has been associated with increased risk of coronary and other atherosclerotic disease[51]. The possibility of a link between variants in *MTHFD1L* and CAD risk is supported by evidence that *MTHFD1L* activity also contributes to plasma homocysteine[52] and that defects in the *MTHFD1L* pathway may increase plasma homocysteine level[48,53].

An intronic SNP in *ADAMTS17* (a disintegrin and metalloproteinase with thrombospondin motifs 17), which showed modest association (rs1994016; $P = 1.1 \times 10^{-4}$) in our primary analysis, showed a much stronger association in the expanded reference group analysis (see below and Supplementary Table 9). Although the specific function of *ADAMTS17* has not been determined, other members of the ADAMTS family have been implicated in vascular extracellular matrix degradation, vascular remodelling and atherosclerosis[54,55].

Crohn's disease (CD). Crohn's disease is a common form of chronic inflammatory bowel disease[56]. The pathogenic mechanisms are poorly understood, but probably involve a dysregulated immune response to commensal intestinal bacteria and possibly defects in mucosal barrier function or bacterial clearance[57]. Genetic predisposition to

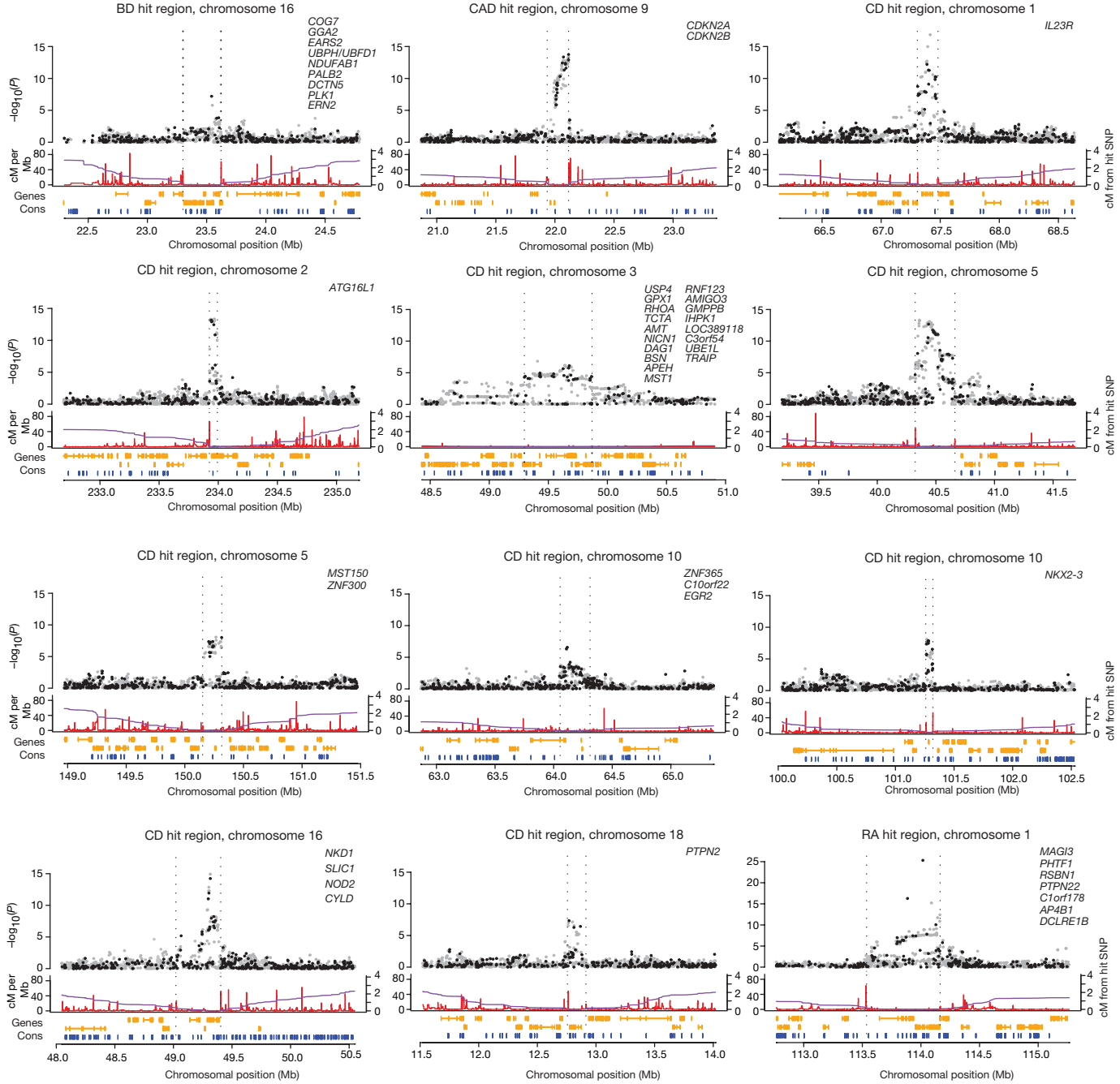
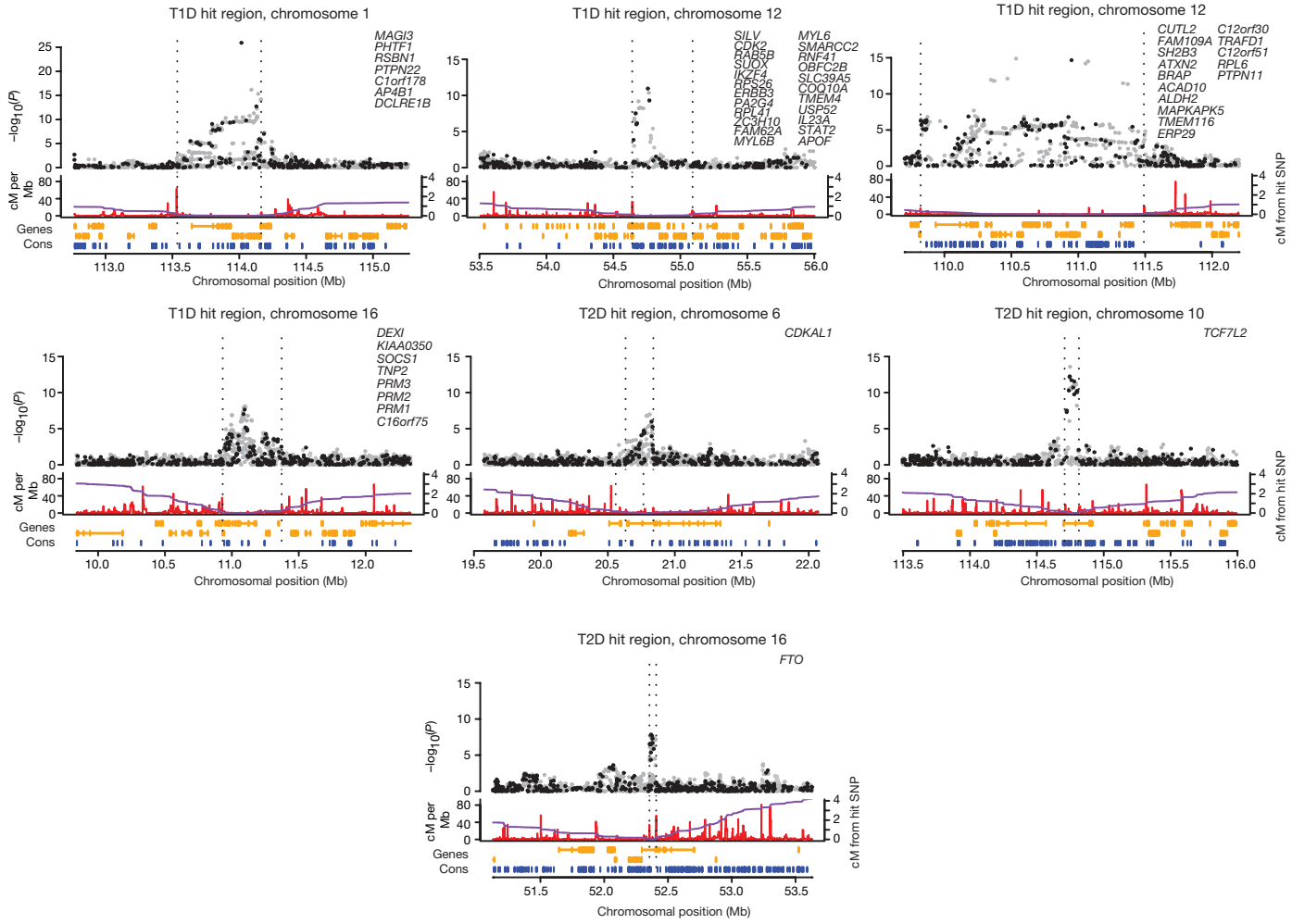


Figure 5 | Regions of the genome showing strong evidence of association. Characteristics of genomic regions 1.25 Mb to either side of 'hit SNPs'—SNPs with lowest *P* values. Region boundaries (vertical dotted lines) were chosen to coincide with locations where test statistics returned to background levels and, where possible, recombination hotspots. Upper panel, $-\log_{10}(P$ values) for the test (trend or genotypic) with the smallest *P* value at the hit SNP. Black points represent SNPs typed in the study, and grey points represent SNPs whose genotypes were imputed. SNPs imputed with higher confidence are shown in darker grey. Middle panel, fine-scale recombination rate (centimorgans per Mb) estimated from Phase II HapMap. The purple line shows the cumulative genetic distance (in cM)

CD is suggested by a λ_s of 17–35 and by twin studies that contrast monozygotic concordance rates of 50% with only 10% in dizygotic pairs[58,59].

A number of CD-susceptibility loci have previously been defined, and all of these generate strong signals in our data (Table 2). In 2001, positional cloning identified *CARD15* (caspase recruitment domain family, member 15; *NOD2*) as the first confirmed CD-susceptibility gene[60,61]. In the present study, this locus is represented by rs17221417 ($P = 9.4 \times 10^{-12}$). A second association, on chromosome 5q31 (ref. 62) has been widely replicated, although the identity of the causative gene is disputed owing to extensive regional linkage disequilibrium[63]. Here, the previously described risk haplotype is tagged by rs6596075 ($P = 5.4 \times 10^{-7}$).

More recent studies have identified four further CD-susceptibility loci, all of which are strongly replicated in the present study. The association between CD and SNPs within *IL23R* (interleukin 23 receptor)[63] is here represented by a cluster of associated SNPs, including rs11805303 ($P = 6.5 \times 10^{-13}$). The strongest signal for CD in the present scan (at rs10210302; $P = 7.1 \times 10^{-14}$) maps to the *ATG16L1* (ATG16 autophagy related 16-like 1) gene and is in strong linkage disequilibrium ($r^2 = 0.97$) with a non-synonymous SNP (T300A, rs2241880) associated with CD in a German non-synonymous SNP scan[64]. The third is a locus at chromosome 10q21 around rs10761659 ($P = 2.7 \times 10^{-7}$) and represents a non-coding intergenic SNP mapping 14-kb telomeric to gene *ZNF365* and 55-kb centromeric to the pseudogene antiquitin-like 4—a



from the hit SNP. Lower panel, known genes, and sequence conservation in 17 vertebrates. Known genes (orange) in the hit region are listed in the upper right part of each plot in chromosomal order, starting at the left edge of the region. The top track shows plus-strand genes and the middle track shows minus-strand genes. Sequence conservation (bottom track) scores are based on the phylogenetic hidden Markov model phastCons. Highly conserved regions (phastCons score ≥600) are shown in blue. Information in middle and lower panels is taken from the UCSC Genome Browser. Positions are in NCBI build-35 coordinates. See Supplementary Information on 'signal plots'.

Table 4 | Regions of the genome showing moderate evidence of association

Collection	Chromosome	Region (Mb)	SNP	Trend *P* value	Genotypic *P* value	$\log_{10}$(BF), additive	$\log_{10}$(BF), general	Risk allele	Minor allele	Heterozygote odds ratio	Homozygote odds ratio	Control MAF	Case MAF
BD	2p25	11.94–12.00	rs4027132	1.31×10^{-05}	9.68×10^{-06}	3.07	2.84	A	G	1.39 (1.19–1.64)	1.51 (1.27–1.79)	0.459	0.414
BD	2q12	104.41–104.58	rs7570682	3.11×10^{-06}	1.64×10^{-05}	3.68	3.23	A	A	1.23 (1.09–1.40)	1.64 (1.28–2.12)	0.214	0.255
BD	2q14	115.63–116.11	rs1375144	2.43×10^{-06}	1.31×10^{-05}	3.80	2.92	A	G	1.32 (1.07–1.63)	1.59 (1.29–1.96)	0.337	0.291
BD	2q37	241.23–241.28	rs2953145	1.11×10^{-05}	6.57×10^{-06}	3.22	3.50	C	G	1.84 (1.31–2.58)	2.14 (1.53–2.98)	0.226	0.189
BD	3p23	32.26–32.33	rs4276227	4.57×10^{-06}	2.62×10^{-05}	3.52	3.04	C	T	1.20 (0.99–1.46)	1.49 (1.23–1.81)	0.371	0.326
BD	3q27	184.29–184.40	rs683395	2.30×10^{-06}	5.11×10^{-06}	3.87	3.73	G	G	1.47 (1.26–1.71)	1.30 (0.69–2.46)	0.080	0.109
BD	6p21	42.82–42.86	rs6458307	3.43×10^{-01}	4.35×10^{-06}	−0.80	2.84	T	T	0.84 (0.75–0.96)	1.39 (1.13–1.69)	0.312	0.321
BD	8p12	34.22–34.61	rs2609653	6.86×10^{-06}	-	3.44	3.21	C	C	1.43 (1.19–1.71)	3.62 (1.26–10.44)	0.052	0.074
BD	9q32	114.31–114.39	rs10982256	8.80×10^{-06}	4.41×10^{-05}	3.23	2.37	T	C	1.26 (1.08–1.47)	1.47 (1.24–1.74)	0.471	0.425
BD	14q22	57.17–57.24	rs10134944	3.21×10^{-06}	6.89×10^{-06}	3.73	3.59	T	T	1.45 (1.24–1.68)	1.32 (0.74–2.33)	0.086	0.115
BD	14q32	103.43–103.62	rs11622475	2.10×10^{-06}	8.14×10^{-06}	3.87	3.24	C	T	1.13 (0.89–1.44)	1.47 (1.17–1.86)	0.300	0.256
BD	16q12	51.36–51.50	rs1344484	1.64×10^{-06}	1.03×10^{-05}	3.94	3.41	T	C	1.24 (1.03–1.48)	1.52 (1.27–1.82)	0.402	0.353
BD	20p13	3.70–3.73	rs3761218	4.43×10^{-05}	6.71×10^{-06}	2.58	3.18	T	C	0.97 (0.81–1.15)	1.31 (1.09–1.57)	0.397	0.356
CAD	1q43	236.77–236.85	rs17672135	1.04×10^{-04}	2.35×10^{-06}	2.36	3.88	T	C	0.70 (0.61–0.81)	1.32 (0.79–2.22)	0.134	0.108
CAD	5q21	99.98–100.11	rs383830	5.72×10^{-06}	1.34×10^{-05}	3.49	3.26	T	A	1.60 (1.16–2.21)	1.92 (1.40–2.63)	0.220	0.182
CAD	6q25	151.34–151.42	rs6922269	6.33×10^{-06}	1.50×10^{-05}	3.38	3.14	A	A	1.17 (1.04–1.32)	1.65 (1.32–2.06)	0.253	0.294
CAD	16q23	81.72–81.79	rs8055236	9.73×10^{-06}	5.60×10^{-06}	3.28	3.59	G	T	1.91 (1.33–2.74)	2.23 (1.56–3.17)	0.198	0.162
CAD	19q12	34.74–34.78	rs7250581	9.12×10^{-06}	2.50×10^{-05}	3.30	2.87	G	A	1.06 (0.79–1.43)	1.40 (1.05–1.86)	0.220	0.182
CAD	22q12	25.01–25.06	rs688034	6.90×10^{-06}	3.75×10^{-06}	3.33	3.15	T	T	1.11 (0.98–1.25)	1.62 (1.34–1.95)	0.310	0.355
CD	1q24	169.53–169.67	rs12037606	1.79×10^{-06}	1.09×10^{-05}	3.89	3.35	A	A	1.22 (1.07–1.40)	1.52 (1.28–1.82)	0.388	0.438
CD	5q23	131.40–131.90	rs6596075	5.40×10^{-07}	3.19×10^{-06}	4.54	4.01	C	G	1.55 (1.00–2.39)	2.06 (1.35–3.14)	0.166	0.127
CD	6p22	20.83–20.85	rs6908425	5.13×10^{-06}	1.10×10^{-05}	3.55	3.38	C	T	1.63 (1.18–2.25)	1.95 (1.43–2.67)	0.230	0.190
CD	6p21	32.79–32.91	rs9469220	8.65×10^{-07}	2.28×10^{-06}	4.19	3.92	A	A	1.14 (0.98–1.32)	1.52 (1.28–1.79)	0.481	0.534
CD	6q23	138.06–138.17	rs7753394	4.42×10^{-06}	2.59×10^{-05}	3.52	2.99	C	C	1.21 (1.04–1.40)	1.48 (1.25–1.76)	0.482	0.531
CD	7q36	147.62–147.70	rs7807268	6.89×10^{-06}	4.42×10^{-06}	3.33	3.58	G	G	1.38 (1.20–1.60)	1.47 (1.24–1.74)	0.462	0.509
CD	10p15	38.52–38.57	rs6601764	2.56×10^{-06}	8.95×10^{-06}	3.74	3.01	C	C	1.16 (1.01–1.33)	1.52 (1.28–1.80)	0.408	0.458
CD	19q13	50.89–51.07	rs8111071	6.14×10^{-06}	1.75×10^{-05}	3.48	3.29	G	G	1.47 (1.25–1.73)	1.28 (0.56–2.88)	0.070	0.096
HT	1q43	235.67–235.79	rs2820037	5.76×10^{-05}	7.66×10^{-07}	2.54	3.99	T	T	1.54 (1.03–2.31)	1.09 (0.74–1.62)	0.141	0.171
HT	8q24	140.17–140.35	rs6997709	7.88×10^{-06}	4.36×10^{-05}	3.32	2.60	G	T	1.20 (0.94–1.52)	1.49 (1.18–1.89)	0.285	0.244
HT	12p12	24.86–24.95	rs7961152	7.39×10^{-06}	3.03×10^{-05}	3.29	2.51	A	A	1.16 (1.01–1.32)	1.47 (1.25–1.74)	0.415	0.461
HT	12q23	100.52–100.58	rs11110912	9.18×10^{-06}	1.94×10^{-05}	3.27	3.11	G	G	1.33 (1.18–1.51)	1.34 (0.96–1.86)	0.165	0.200
HT	13q21	66.90–67.04	rs1937506	9.23×10^{-06}	4.53×10^{-05}	3.25	2.85	G	A	1.33 (1.04–1.69)	1.60 (1.26–2.02)	0.289	0.248
HT	15q26	94.60–94.67	rs2398162	7.85×10^{-06}	5.67×10^{-06}	3.33	3.40	A	G	0.97 (0.76–1.25)	1.31 (1.03–1.67)	0.258	0.218
RA	1p36	2.44–2.77	rs6684865	5.37×10^{-06}	3.14×10^{-05}	3.47	2.97	G	A	1.27 (1.02–1.56)	1.54 (1.25–1.90)	0.338	0.294
RA	1p31	80.16–80.36	rs11162922	1.80×10^{-06}	-	4.11	3.80	A	G	1.27 (0.41–4.01)	2.00 (0.64–6.20)	0.072	0.048
RA	4p15	24.99–25.13	rs3816587	7.65×10^{-03}	9.25×10^{-06}	0.50	2.64	C	C	0.91 (0.80–1.04)	1.35 (1.14–1.59)	0.406	0.434
RA	6q23	138.00–138.06	rs6920220	4.99×10^{-06}	1.58×10^{-05}	3.49	3.17	A	A	1.20 (1.06–1.36)	1.72 (1.33–2.22)	0.223	0.263
RA	7q32	130.80–130.84	rs11761231	1.74×10^{-06}	2.65×10^{-06}	3.92	3.42	C	T	1.44 (1.19–1.75)	1.64 (1.35–1.99)	0.375	0.327
RA	10p15	6.07–6.16	rs2104286	7.02×10^{-06}	2.52×10^{-05}	3.37	2.57	T	C	1.41 (1.10–1.81)	1.68 (1.31–2.14)	0.286	0.244
RA	13q12	19.845–19.855	rs9550642	8.44×10^{-06}	3.90×10^{-05}	3.35	3.02	A	A	1.34 (1.15–1.56)	2.23 (1.21–4.13)	0.084	0.112
RA	21q22	41.430–41.465	rs2837960	3.45×10^{-02}	1.68×10^{-06}	0.05	2.70	G	G	0.95 (0.83–1.08)	2.30 (1.64–3.23)	0.171	0.188
RA	22q13	35.870–35.885	rs743777	7.92×10^{-06}	1.15×10^{-06}	3.29	3.52	G	G	1.09 (0.97–1.24)	1.72 (1.40–2.11)	0.292	0.336
T1D	1q42	221.92–222.17	rs2639703	8.46×10^{-06}	1.74×10^{-05}	3.25	3.06	C	C	1.15 (1.02–1.30)	1.61 (1.31–1.99)	0.276	0.318
T1D	4q27	123.02–123.92	rs17388568	5.01×10^{-07}	3.27×10^{-06}	4.42	3.89	A	A	1.26 (1.11–1.42)	1.58 (1.27–1.95)	0.260	0.307
T1D	5q14	86.20–86.50	rs2544677	8.23×10^{-06}	4.43×10^{-05}	3.32	2.70	C	G	1.34 (1.00–1.79)	1.65 (1.24–2.18)	0.242	0.204
T1D	5q31	132.64–132.67	rs17166496	6.06×10^{-01}	5.20×10^{-06}	−0.97	3.25	C	G	0.77 (0.68–0.87)	1.09 (0.92–1.29)	0.391	0.386
T1D	10p15	6.07–6.18	rs2104286	7.96×10^{-06}	4.32×10^{-05}	3.31	2.88	T	C	1.30 (1.02–1.65)	1.57 (1.25–1.99)	0.286	0.245
T1D	12p13	9.71–9.80	rs11052552	1.02×10^{-04}	7.24×10^{-07}	2.22	3.80	G	T	1.49 (1.28–1.73)	1.43 (1.21–1.69)	0.486	0.446
T1D	18p11	12.76–12.91	rs2542151	1.89×10^{-06}	1.16×10^{-05}	3.91	3.52	G	G	1.30 (1.15–1.47)	1.62 (1.17–2.24)	0.163	0.201
T2D	1p31	66.04–66.36	rs4655595	2.68×10^{-06}	1.33×10^{-05}	3.81	3.47	G	G	1.37 (1.17–1.59)	2.33 (1.23–4.42)	0.080	0.108
T2D	2q24	160.90–161.17	rs6718526	2.40×10^{-06}	1.16×10^{-05}	3.86	3.35	C	T	1.49 (1.05–2.11)	1.86 (1.32–2.63)	0.209	0.171
T2D	3p14	55.24–55.32	rs358806	4.77×10^{-01}	3.05×10^{-06}	−0.83	2.72	A	A	0.86 (0.75–0.97)	1.78 (1.34–2.36)	0.198	0.204
T2D	4q27	122.92–123.02	rs7659604	2.1×10^{-02}	9.42×10^{-06}	0.13	2.74	T	T	1.35 (1.19–1.54)	1.09 (0.91–1.30)	0.380	0.403
T2D	10q11	43.43–43.63	rs9326506	7.78×10^{-06}	2.99×10^{-05}	3.27	2.92	C	C	1.28 (1.11–1.48)	1.46 (1.24–1.72)	0.492	0.538
T2D	12q13	49.50–49.87	rs12304921	5.37×10^{-02}	7.07×10^{-06}	−0.09	2.68	G	G	2.50 (1.53–4.09)	1.94 (1.20–3.15)	0.145	0.159
T2D	12q15	69.58–69.96	rs1495377	1.31×10^{-06}	6.52×10^{-06}	4.01	3.15	G	G	1.28 (1.11–1.49)	1.51 (1.28–1.78)	0.497	0.547
T2D	15q24	72.24–72.50	rs2930291	7.72×10^{-06}	4.40×10^{-05}	3.30	2.42	G	A	1.25 (1.04–1.51)	1.50 (1.24–1.82)	0.377	0.332
T2D	15q25	78.12–78.36	rs2903265	9.57×10^{-06}	4.98×10^{-05}	3.24	2.53	G	A	1.18 (0.93–1.49)	1.47 (1.17–1.86)	0.284	0.243

Regions with at least one SNP with a *P* value of greater than 5×10^{-7} and less than 1×10^{-5} for either the trend or the genotypic test. Columns as for Table 3. Cluster plots for each SNP have been inspected visually. Positions are in NCBI build-35 coordinates. Genotypic *P* values were not calculated for SNPs with the lowest MAFs owing to low numbers of rare-allele homozygotes and sensitivity to genotype calling errors.

recently detected signal[65]. Finally, strong association with a cluster of SNPs around rs17234657 ($P = 2.1\times10^{-13}$) within a 1.2 Mb gene desert on chromosome 5p13.1, recapitulates the finding of a recent GWA study[66].

The current study identifies four further new strong association signals in CD, located on chromosomes 3p21, 5q33, 10q24 and 18p11 (Table 3; Fig. 5). Successful replication for all four loci is reported elsewhere[23].

The first of these includes several SNPs around *IRGM* (immunity-related guanosine triphosphatase; the human homologue of the mouse *Irgm/Lrg47*), the strongest signal being at rs1000113 ($P = 5.1\times10^{-8}$). *IRGM* encodes a GTP-binding protein which induces autophagy and is involved in elimination of intracellular bacteria, including *Mycobacterium tuberculosis*[67]. Reduced function and/or activity of this gene would be expected to lead to persistence of intracellular bacteria, consistent with existing models of CD pathogenesis[57] and the recent *ATG16L1* association[64] (see above).

The second novel CD association is seen at rs9858542 ($P = 7.7\times10^{-7}$), a synonymous coding SNP within the *BSN* (bassoon) gene on chromosome 3p21. *BSN* is thought to encode a scaffold protein expressed in brain and involved in neurotransmitter release; a more plausible regional candidate is *MST1* (macrophage

stimulating 1), which encodes a protein influencing motile activity and phagocytosis by resident peritoneal macrophages[68].

The third novel association involves a cluster of SNPs around rs10883365 ($P = 1.4 \times 10^{-8}$) on chromosome 10q24.2. The most credible candidate here is the *NKX2-3* (NK2 transcription factor related, locus 3) gene, a member of the NKX family of homeodomain-containing transcription factors. Targeted disruption of the murine homologue of *NKX2-3* results in defective development of the intestine and secondary lymphoid organs[69]. Abnormal expression of *NKX2-3* may alter gut migration of antigen-responsive lymphocytes and influence the intestinal inflammatory response.

The final novel association, at rs2542151 ($P = 4.6 \times 10^{-8}$) maps 5.5-kb upstream of *PTPN2* (protein tyrosine phosphatase, non-receptor type 2) on chromosome 18p11. *PTPN2* encodes the T cell protein tyrosine phosphatase TCPTP, a key negative regulator of inflammatory responses. The same locus also shows strong association with T1D susceptibility (trend test $P = 1.9 \times 10^{-6}$) and a consistent, though weaker, association with RA ($P = 1.9 \times 10^{-2}$), supporting the existence of overlapping pathways in the pathogenesis of very distinct inflammatory phenotypes (combined trend test *P* value for all three diseases $= 9 \times 10^{-8}$) (Table 3; ref. 10).

Several further loci generating less strong evidence for association are of interest on the basis of their biological candidacy (Table 4). For example, rs9469220 ($P = 8.7 \times 10^{-7}$) mapping to the human leukocyte antigen (HLA) system class II region was detected in the 'second tier' of associations (Table 4). This suggests a significant contribution of HLA to CD-susceptibility, though less marked than seen in classical autoimmune conditions such as RA and T1D. Another interesting candidate flagged in Table 4 is *TNFAIP3* (TNFα induced protein 3), the closest gene to rs7753394 on chromosome 6q23. The protein product inhibits TNFα-induced NFκB-dependent gene expression by interfering with RIP- or TRAF-2-mediated transactivation signals—hence interacting with the same pathway as *CARD15* (*NOD2*). Markers with lower levels of significance include rs6478108 ($P = 9.0 \times 10^{-5}$) within *TNFSF15* (tumour necrosis factor super family, member 15), previously reported associated with CD[70]; and rs3816769 ($P = 3.1 \times 10^{-5}$) which maps within *STAT3* (signal transducers and activator of transcription, member 3). On the X chromosome rs2807261 ($P = 1.3 \times 10^{-7}$) maps 50-kb from the gene *CD40LG* (CD40 ligand—previously known as TNF superfamily, member 5), implicated in the regulation of B-cell proliferation, adhesion and immunoglobulin class switching[71]. As described in the section on T1D, a modest association between CD and SNPs in the vicinity of the *PTPN11* gene on chromosome 12q24 ($P = 1.5 \times 10^{-3}$) probably reflects a locus influencing general autoimmune predisposition.

An emerging theme from molecular genetic studies of CD is the importance of defects in autophagy and the processing of phagocytosed bacteria. A number of other specific components within innate and adaptive immune pathways are also highlighted.

Hypertension (HT). Hypertension refers to a clinically significant increase in blood pressure and constitutes an important risk factor for cardiovascular disease (http://www.who.int/whr/2002/en/; ref. 72). Lifestyle exposures that elevate blood pressure, including sodium intake, alcohol and excess weight[73] are well-described risk factors. Genetic factors are also important[74,75]. Estimates of λ_s are approximately 2.5–3.5.

Experimental models have highlighted a number of quantitative trait loci but these have yet to translate into insights into human hypertension[76]. Linkage studies are consistent with susceptibility genes of modest effect size[77] and well-replicated findings have yet to emerge from association approaches.

None of the variants previously associated with HT showed evidence for association in our study although we note that some, such as promoter of the *WNK1* (WNK lysine deficient protein kinase 1) gene[78,79], are not well tagged by the Affymetrix chip.

For HT there were no SNPs with significance below 5×10^{-7} (Table 3) but the number and distribution of association signals in the range 10^{-4} to 10^{-7} was similar to that of the other diseases studied (Table 4 and Supplementary Table 7). There are several possible explanations. First, HT may have fewer common risk alleles of larger effect sizes than some of the other complex phenotypes. If so, then identification of susceptibility variants for HT is likely to be reliant on the synthesis of findings from multiple large-scale studies. Second, the present study may have failed to detect genuine common susceptibility variants of large effect size because they happened to be poorly tagged by the set of SNPs genotyped in the current study. If so, further rounds of genotyping using resources that offer increased density (or complementary SNP sets), and/or improved analytical methods (for example, imputation-based) should facilitate their discovery. Third, study of HT may be more susceptible than other phenotypes to the diluting effects of misclassification bias due to the presence of hypertensive individuals within the control samples. If so, power can be improved in future studies by use of controls specifically screened to exclude individuals with elevated blood pressure.

The most strongly associated SNPs (Table 4) do not identify genes from physiological systems previously implicated by clinical or genetic studies in hypertension. The strongest signal overall is with rs2820037 on 1q43 (genotypic test, $P = 7.7 \times 10^{-7}$). The closest genes are *RYR2* (encoding the ryanodine receptor 2), mutations in which are associated with stress-induced polymorphic ventricular tachycardia and arrhythmogenic right ventricular dysplasia[80,81]; *CHRM3*, encoding the cholinergic receptor muscarinic 3, a member of the G protein-coupled receptor family[32]; and *ZP4,* the product of which is zona pellucida glycoprotein 4[81]. The strong association signals on the X chromosome using an expanded reference group (see below and Supplementary Table 9) are of substantial interest but they do not identify known genes of obvious relevance to HT.

Rheumatoid arthritis (RA). Rheumatoid arthritis is a chronic inflammatory disease characterized by destruction of the synovial joints resulting in severe disability, particularly in patients who remain refractory to available therapies[82]. Susceptibility to, and severity of, RA are determined by both genetic and environmental factors, with λ_s estimates ranging from 5–10 (ref. 83).

An association between RA and alleles of the *HLA-DRB1* locus has long been established[84]. Despite extensive linkage[85–87] and association studies, only one other RA susceptibility locus has been convincingly identified in Caucasians. In common with several autoimmune diseases including T1D, carriage of the T allele of the rs2476601 SNP in the *PTPN22* (protein tyrosine phosphatase, non-receptor type 22) gene has been reproducibly associated with RA, conferring a genetic relative risk of approximately 1.8 (refs 88, 89). These known associations with *HLA-DRB1* and *PTPN22* explain around 50% of the familial aggregation of RA.

Both these previous associations emerge strongly here (Table 2). The most associated marker within *PTPN22* (rs6679677: chromosome 1p13) is perfectly correlated (HapMap CEU data $r^2 = 1$) with the functionally relevant SNP (rs2476601) described previously, and the effect size is consistent with previous estimates[89]. Amongst other putative RA susceptibility genes, two SNPs mapping to *CTLA-4* (cytotoxic T-lymphocyte associated 4) rs3087243 and rs11571300 were only nominally significant ($P = 0.085$ and $P = 0.034$, respectively) (Supplementary Table 10).

RA was the sole disease for which the sex-differentiated analysis generated a strong signal due to different genetic effects in males and females. The SNP rs11761231 (chromosome 7) generates a *P* value of 3.9×10^{-7} for the 2-degrees of freedom (d.f.) sex-differentiated test which combines trend tests in males and females (Table 3). (The trend test ignoring the sex of the individuals has a *P* value of 1.7×10^{-6}.) This genotype has no effect on disease status in males, but a strong apparently additive effect in females (*P* value in a logistic regression model with additive log-odds is 0.68 in males and 6.8×10^{-8} in females, additive OR for females 1.32), and may represent one of the first sex-differentiated effects in human diseases. Cluster plots for this SNP seem good, but it is surrounded by

recombination hotspots and has no other SNPs on the Affymetrix chip with $r^2 > 0.1$ (Supplementary Fig. 11). Some caution is therefore required, but this represents a potentially interesting finding which warrants further investigation, particularly given the sex-related prevalence difference characteristic of this condition.

None of the 9 SNPs with nominal *P* values in the range 10^{-5} to 5×10^{-7} (Table 4) map to loci previously associated with RA. Of particular interest is the association of SNPs mapping close to both the alpha and beta chains of the IL2 receptor (rs2104286 in the case of *IL2RA;* rs743777 and *IL2RB*). The IL2 receptor mediates IL2 stimulation of T lymphocytes and is thereby thought to have an important role in preventing autoimmunity. A rare 4-base-pair deletion of *IL2RA* has been associated with development of severe autoimmune disease[90], and there is evidence (from previous data[91], and from this study and its follow-up) that SNPs within the *IL2RA* gene region are associated with T1D (see also T1D section).

Several of the SNPs with nominal significance in the range 10^{-4} to 10^{-5} (Supplementary Table 7) map to genes with plausible biological relevance. Examples include SNPs within genes implicated in the TNF pathway (for example, rs2771369 in *TNFAIP2* (tumour necrosis factor, alpha-induced protein 2)) or in the regulation of T-cell function (rs854350 in *GZMB* (granzyme B) and rs4750316 in *PRKCQ* (protein kinase C, theta)). The association with rs10786617 in *KAZALD1* (Kazal-type serine protease inhibitor domain-containing protein 1 precursor), a gene whose product is known to have a role in bone regeneration after injury, may be relevant to the development of bone erosions in RA.

RA and T1D were already known to have two disease susceptibility genes in common: at the MHC, and at *PTPN22*. As detailed elsewhere, our study provides data indicating that this list can be extended to include variants around *IL2RA* (chromosome 10p15), *PTPN2* (chromosome 18p11) and the chromosome 12q24 region (Supplementary Table 11), all apparently novel in RA.

Type 1 diabetes (T1D). Type 1 diabetes is a chronic autoimmune disorder with onset usually in childhood[92]. The λ_s for T1D is ~15 and twin data suggest that over 85% of the phenotypic variance is due to genetic factors[93]. There are six genes/regions for which there is strong pre-existing statistical support for a role in T1D-susceptibility: these are the major histocompatibility complex (MHC), the genes encoding insulin, CTLA-4 (cytotoxic T-lymphocyte associated 4) and PTPN22 (protein tyrosine phosphatase, non-receptor type 22), and the regions around the interleukin 2 receptor alpha (*IL2RA/CD25*) and interferon-induced helicase 1 genes (*IFIH1*/MDA5)[94]. However, these signals can explain only part of the familial aggregation of T1D. Five of these previously identified associations were detected in this scan ($P \leq 0.001$) (Table 2 and Supplementary Table 10), the exception being the *INS* gene discussed above.

In this study, single-point analyses revealed three novel regions (on chromosomes 12q13, 12q24 and 16p13) showing strong evidence of association ($P < 5 \times 10^{-7}$; Table 3). Four further regions attained similar levels of significance either through multilocus analyses (chromosomes 4q27 and 12p13: Table 3, Supplementary Fig. 12), or through the combined analysis of autoimmune cases (chromosomes 18p11 and the 10p15 CD25 region: Table 3, Supplementary Fig. 13). The associations with T1D for chromosomes 12q13, 12q24, 16p13 and 18p11 have been confirmed in independent and multiple populations[10].

The two signals on chromosome 12 (at 12q13 and 12q24) map to regions of extensive linkage disequilibrium covering more than ten genes (Fig. 5). Several of these represent functional candidates because of their presumed roles in immune signalling, considered to be a major feature of T1D-susceptibility. These include *ERBB3* (receptor tyrosine-protein kinase erbB-3 precursor) at 12q13 and *SH2B3/LNK* (SH2B adaptor protein 3), *TRAFD1* (TRAF-type zinc finger domain containing 1) and *PTPN11* (protein tyrosine phosphatase, non-receptor type 11) at 12q24. For these signal regions in particular, extensive resequencing, further genotyping and targeted functional studies will be essential steps in identifying which gene, or genes, are causal[95]. Of those listed, *PTPN11* is a particularly attractive candidate given a major role in insulin and immune signalling[96]. It is also a member of the same family of regulatory phosphatases as *PTPN22*, already established as an important susceptibility gene for T1D and other autoimmune diseases[94,97]. Indeed, the 12q24 variant most associated with T1D also features in both the CD and RA scans, generating a combined signal for all autoimmune cases of 9.3×10^{-10} (Supplementary Table 11).

In contrast, available annotations suggest that the 16p13 region contains only two genes of unknown function, *KIAA0350* and dexamethasone-induced transcript (Fig. 5). Also, the region of association identified on 18p11 (Supplementary Fig. 14), which seems to confer susceptibility to all three autoimmune conditions studied (combined trend test $P = 9 \times 10^{-8}$, $P = 4.6 \times 10^{-8}$ for CD, 1.9×10^{-2} for RA, and 1.9×10^{-6} for T1D: Supplementary Table 11), maps to a single gene, *PTPN2* (protein tyrosine phosphatase, non-receptor type 2), a member of the same family as *PTPN22* and *PTPN11* and involved in immune regulation[96].

Our scan found associations with SNPs within the chromosome 10p15 region containing CD25, encoding the high-affinity receptor for IL-2. This is consistent with a previous report of associations of this region with T1D[91]. The *CD25* region has previously been shown to be associated with Graves' disease[98] and the present study also provides evidence of association with RA (combined trend test $P = 5 \times 10^{-8}$, $P = \sim 7 \times 10^{-6}$ for RA and T1D separately, Supplementary Table 11). This finding has clear biological connections to the evidence of association between T1D and a region of 4q27 revealed by the multilocus analysis (Supplementary Table 12, Supplementary Fig. 12). This region contains the genes encoding both IL-2 and IL-21. Together with studies in the NOD (nonobese diabetic) mouse model of T1D, which have shown that a major non-MHC locus (*Idd3*) reflects regulatory variation of the *Il2* gene[99], our results point to the primary importance of the IL-2 pathway in T1D and other autoimmune diseases.

One further region deserves comment. In the multilocus analysis, there was increased support for a region on chromosome 12p13 containing several candidate genes, including *CD69* (CD69 antigen (p60, early T-cell activation antigen)) and multiple *CLEC* (C-type lectin domain family) genes. In contrast to the chromosome 4 region where the effect of imputation is to tip an already-strong signal (5.01×10^{-7} for typed rs17388568, trend test) over the arbitrary threshold of 5×10^{-7}, the 12p13 locus involves a more marked change between imputed and actual (7.2×10^{-7} for rs11052552, general test). Replication studies of this imputed SNP to date have produced equivocal results (for details see ref. 10).

Type 2 diabetes (T2D). Type 2 diabetes is a chronic metabolic disorder typically first diagnosed in the middle to late adult years[100]. Strongly associated with obesity, the condition features defects in both the secretion and peripheral actions of insulin[101]. The appreciable familial aggregation of T2D (an estimated λs of ~3.0 in European individuals)[73] reflects both shared family environment and genetic predisposition. Heritability values vary widely with most estimates between 30 and 70%[101].

To date, robust, widely replicated associations in non-isolate populations are limited to variants in three genes: *PPARG* (encoding the peroxisomal proliferative activated receptor gamma; P12A[102]), *KCNJ11* (the inwardly-rectifying Kir6.2 component of the pancreatic beta-cell KATP channel; E23K[103]) and *TCF7L2* (transcription factor 7-like 2; rs7903146 (refs 104, 105)).

All three of these signals are detected here with effect-sizes consistent with previous reports (Table 2). A cluster of SNPs on chromosome 10q, within *TCF7L2,* represented by rs4506565 (trend test, OR 1.36, $P = 5.7 \times 10^{-13}$) generates the strongest association signal for T2D (Table 3, Fig. 5). Rs4506565 is in tight linkage disequilibrium (r^2 of 0.92 in the CEU component of HapMap) with rs7903146, the variant with the strongest aetiological claims[104,106]. In fact, our

imputation analysis confirms that rs7903146, though unrepresented on the chip, is responsible for the strongest association effect in this region (Fig. 5). TCF7L2 acts within the WNT-signalling pathway, and effects on diabetes risk seem to be mediated predominantly through beta-cell dysfunction[107].

As expected, given existing effect-size estimates, the signals associated with variants within the other established T2D-susceptibility genes, *KCNJ11* (rs5215, r^2 of 0.9 with rs5219, E23K) and *PPARG* (rs17036328, r^2 of 1 with rs1801282, P12A) are less dramatic (trend test, OR 1.15 and 1.23 respectively, both $P = \sim 0.001$). These examples illustrate how genuine disease-susceptibility variants can generate association signals which would not attract immediate attention for follow-up in the genomewide context.

Apart from *TCF7L2,* the scan reveals two signals for T2D with *P* values less than 5×10^{-7} (Table 3, Fig. 5). The first of these maps within the *FTO* (fat-mass and obesity-associated) gene on chromosome 16q. Several adjacent SNPs (including rs9939609, rs7193144 and rs8050136) generate signals characterized by a per-allele OR for T2D of ~1.25 and a risk-allele frequency of ~40% in controls. As recently described in follow-up studies prompted by this finding, the effect of these variants on T2D-risk has been replicated and is mediated entirely by their marked effect on adiposity[24].

The third association signal (chromosome 6p22) features a cluster of highly associated SNPs (including rs9465871) with risk-allele frequencies between 18 and 35%, mapping to intron 5 of the *CDKAL1* (CDK5 regulatory subunit associated protein 1-like 1) gene. Although the function of CDKAL1 is not known, it shares homology at the protein domain level with CDK5 regulatory subunit associated protein 1 (CDK5RAP1). CDK5RAP1 is known to inhibit the activation of CDK5, a cyclin-dependent kinase which has been implicated in the maintenance of normal beta-cell function[108]. Our own follow-up studies, and scans by other groups have shown strong replication of this finding[19–22]. The effect of this variant on T2D-risk shows significant departures from additivity (Supplementary Table 8).

One notable inclusion amongst the variants with more modest association signals is a cluster of SNPs on chromosome 10 including rs10748582 and rs7923866, which generate trend test *P* values between 10^{-4} and 10^{-5}. This cluster maps in the vicinity of the *HHEX* (homeobox, hematopoietically expressed) and *IDE* (insulin-degrading enzyme) genes, in a region recently highlighted in a GWA scan for T2D performed in 1363 subjects of French origin[109]. The SNPs showing association in our data are proxies for those reported in the French study and generate similar effect-size estimates for T2D.

Of the three other regions highlighted by the French scan[109], none can be confirmed by our data. The SNP in *SLC30A8* associated with T2D in the French report (rs13266634) is poorly correlated with SNPs on the Affymetrix chip ($r^2 < 0.01$), and extensive recombination events in the region limit the value of data-imputation methods. Coverage of the *LOC387761* and *EXT2* signals is considerably better, but, for these, neither genotyped nor imputed SNPs show evidence for association with T2D.

WTCCC data contributed to identification of two additional robustly replicating T2D signals, mapping to the *IGF2BP2* gene and *CDKN2A/CDKN2B* regions[19,21,22], although neither generated impressive *P* values on the primary scan analysis (neither single-point *P* was $<10^{-4}$). The latter signal maps to the same region as the CAD signal on chromosome 9 though different SNPs are involved. The other SNPs in Table 4 do not map to genes or regions previously implicated in T2D pathogenesis, and replication efforts to date have not identified any confirmed signals[19].

Expanded reference group analyses. For a fixed number of cases, power of a case-control study can be increased by enlarging the reference group. Our main analyses used a control:case ratio of 1.5:1 for each disease. The availability of the other 6 disease data sets gave us the opportunity to expand the reference group up to a ratio of ~7.5:1, with potential reciprocal benefits for the analysis of each disease. For BD and T2D the expanded reference group comprised the 58C and UKBS controls supplemented by the other 6 disease sets; for CAD and HT this expanded reference group was reduced to exclude HT and CAD respectively; for CD, RA and T1D, the reference group was augmented only by the cases from the non-autoimmune diseases.

The utility of the expanded reference group approach was demonstrated by increased evidence for association at most of the loci that received strongest support from our primary analysis, including many of the signals at loci known to show robust association in T1D, T2D and CD (Supplementary Table 9). Additionally, this analysis elevated several loci with modest levels of statistical significance in the primary analysis, to the top tier of statistical significance ($P < 5 \times 10^{-7}$).

Our data indicate that this approach may be a useful adjunct to conventional analysis and that loci identified as highly significant should be considered for follow up. There are two important caveats. First, susceptibility genes that influence both the test disease and one or more of the diseases included in the reference group will cause loss of power. Second, a 'mirror-image' effect could occur whereby a strong association within the expanded reference sample (for example, HLA in autoimmune diseases) causes spurious association with the opposite allele in the test disease. Thus, a positive association using an expanded reference group must be interpreted within the context of association findings in the diseases included within the reference group.

Disease models. It is of interest to consider which statistical models best describe the data at and between loci that are strongly associated with disease status. Biological interpretation of these statistical models is not straightforward but they can help in choosing more powerful statistical tools for detecting associations.

First, consider separately each of the 19 non-MHC SNPs showing strong evidence for association on either the trend or genotypic test in Table 3. For four of these 19, the *P* value on the 2-d.f. genotypic test was smaller than that on the 1-d.f. trend test (Table 3). When comparing disease models, these were also the four SNPs with evidence for departure from a simple model in which odds of disease increase multiplicatively with the number of copies of the risk allele (Supplementary Table 8). This supports our view that the genotypic test should be carried out in addition to the trend test, although should perhaps be viewed more cautiously for two reasons: it is more susceptible to genotyping errors; and (on the basis of our findings) experience does not favour strong dominance effects.

A separate question relates to the best models for the way in which different loci combine to affect susceptibility to a disease, and as a consequence on the extent to which methods explicitly allowing interactions between loci should be employed to detect associations[110]. None of the analyses reported here includes such interactions, so we are not well placed to address the general question. Nonetheless, within each collection with multiple associated regions (CD, T1D and T2D) we considered all pairs of non-MHC SNPs in Table 3 and looked for a departure from the model in which the two loci combine to increase log-odds in an additive fashion. We found suggestive evidence of a departure from multilocus additivity between rs1000113 and rs10761659 in CD (unadjusted *P* value = 0.002) and between rs9465871 and rs4506565 in T2D (unadjusted *P* value = 0.004). Further investigation of this question, preferably on unbiased sets of disease loci found through the application of single locus and interaction-based approaches, would seem warranted.

Discussion

We have studied seven common familial diseases by genome-wide association analysis in 16,179 individuals. Our findings inform understanding of the genetic basis of the diseases concerned and provide methodological insights relevant to the pursuit of GWA studies in general.

A simple but important observation is that GWA analysis provides a highly effective approach for exploring the genetic underpinnings of common familial diseases. Our yield of novel, highly significant

association findings is comparable to, or exceeds, the number of those hitherto-generated by candidate gene or positional cloning efforts. For many of the compelling signals, replication has already been obtained, including regions on chromosomes 3p21, 5q33, 10q24 and 18p11 for CD[23], 12q13, 12q24, 16p13 and 18p11 in T1D[10] and 6p22 and 16q12 in T2D[19–22,24]. For others, replication is required to establish a definitive relationship with disease. Additional findings of particular interest include the identification of several loci that seem to influence susceptibility to multiple autoimmune diseases, and the suggestion of a novel locus for RA which shows sex-specific effects.

Our study enables us to make several general recommendations relevant to GWA studies. The first relates to the importance of careful quality control. In such large data sets, small systematic differences can readily produce effects capable of obscuring the true associations being sought[111,112]. We implemented extensive quality control checks to minimize differences in sample DNA concentration, quality and handling procedures and combined a new genotype-calling algorithm (CHIAMO) with a set of filtering heuristics to select SNPs for further analysis. Given that infallible detection of incorrect genotype calls is not yet possible, the criteria used for SNP exclusion need to strike a compromise between stringency (which may discard true signals or generate spurious positives through differential missingness) and leniency (with the danger that true signals are swamped by spurious findings due to poor genotype calling). As such, systematic visual inspection of cluster plots for SNPs of interest remains an integral part of the quality control process.

The potential for population structure to undermine inferences in case-control association studies has long been debated[113] but limited empirical data have been available to assess the issue. Our study highlighted several loci, some known and some new, which demonstrate substantial geographical variation in allele frequencies across Britain (Table 1), most probably due to natural selection in ancestral populations. Outside these loci, the effects of population structure are relatively minor, and do not represent a major source of confounding, provided that individuals with appreciable non-European ancestry are excluded. Although these conclusions may not generalize to studies in other locations, this finding reinforces the logistical and economic benefits of the case-control design over alternatives (such as family-based association studies).

Our study allowed us to address another important methodological issue: the adequacy, or otherwise, of using a common set of controls, rather than a sample recruited explicitly for use with a defined disease sample. It is often assumed that failure to match cases and controls for socio-demographic variables will lead to substantial inflation of the type I error rate. Our study demonstrates that, within the context of large-scale genetic association studies, for British populations at least, this concern has been overstated. A related argument against use of population controls relates to the perceived impact of misclassification bias when a proportion of controls meet the criteria used to define cases. However, the consequent loss of power is modest unless the trait of interest is very common[6]. Given the above, the present study provides a compelling case for both the suitability and efficiency of the common control design in Britain and warrants its serious consideration elsewhere. Further benefits can be expected from use of this common control genotype data set in future GWA studies in Britain. Finally, in failing to detect significant differences in performance between the epidemiological sample (58C) and that derived from blood donors (UKBS), we validate the use of the latter samples for cost-effective, large-scale control DNA provision.

In terms of general biological insights, the most profound relate to inferences about the allelic architecture of common traits. The novel variants we have uncovered are characterized by modest effect size (that is, per-allele ORs between 1.2 and 1.5) and even these estimates are likely to be inflated[114]. We identified no additional common variants of very large effect (akin to HLA in T1D: Supplementary Fig. 15). The observed distribution of effect sizes is consistent with models based on theoretical considerations and empirical data from animal models[87,115,116] that suggest that, for any given trait, there will be few (if any) large effects, a handful of modest effects and a substantial number of genes generating small or very small increases in disease risk.

There are several important corollaries. Notwithstanding the incomplete coverage afforded by the genotyping reagents employed, most of the susceptibility effects yet to be uncovered for these diseases (at least those attributable to, or tagged by, common SNPs) are likely to have effects of similar or smaller magnitude to those we have highlighted. Beyond the signals with the strongest evidence for association, most of which are likely to be real (and many of which have already been confirmed), there will be many additional susceptibility variants for which the WTCCC provides some evidence, but for which extensive replication will be required to establish validity. *PPARG* and *KCNJ11* provide examples of proven susceptibility genes (for T2D) that generated only modest evidence for association within the WTCCC, and which would only have been revealed by such replication efforts. Given the likely preponderance of susceptibility variants of small effect, the potential for identifying further loci is limited only by the clinical resources available for replication (assuming suitable study design, accurate genotyping and appropriate analysis and inference). Provided the attribution of a causal relationship with the trait of interest is robust, even variants of very small effect can offer fundamental biological insights.

The patterns of allelic architecture uncovered mean that replication efforts will need to feature comparably large sample sizes: even if one accepts more relaxed significance thresholds given the prior evidence, one has to consider the inflation in effect-size estimates in the primary study. Caution is required in reaching negative conclusions on the basis of a single failed attempt at replication, or any set of replication attempts that are inadequately powered.

One of our major design considerations was sample size. We set out to include samples larger than those previously examined for genome-wide association, and our results suggest that such large sample sizes were necessary. Even with 2,000 cases and 3,000 controls,

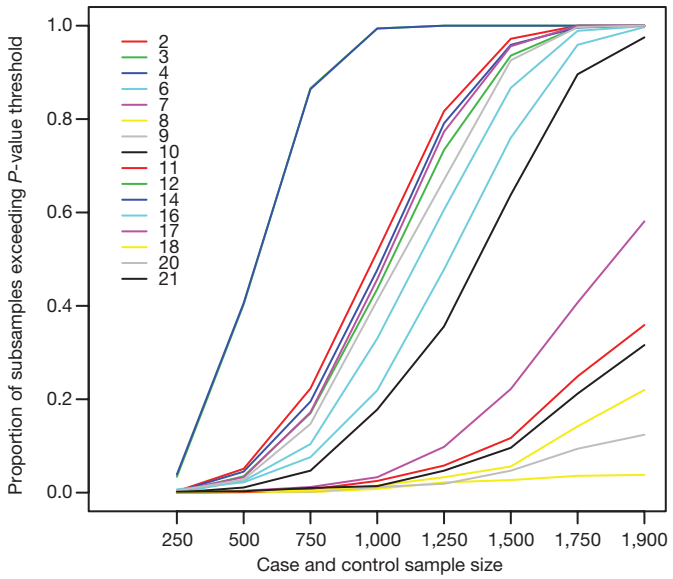


Figure 6 | Strong associations in subsamples of our data. For the 16 SNPs in Table 3 (outside the MHC) with *P* values for the trend test below 5×10^{-7}, we randomly generated 1,000 subsets of our full data set corresponding to case-control studies with different numbers of cases, and the same number of controls (*x* axis). The *y* axis gives the proportion of subsamples of a given size in which that SNP achieved a *P* value for the trend test below 5×10^{-7}. SNPs are numbered according to the row in which they occur in Table 3 (so that, for example, the CAD hit is numbered 2, and the *TCF7L2* hit on chromosome 10 for T2D is numbered 20).

adequate power is restricted to common variants of relatively large effect (see Supplementary Table 2). We carried out an experiment to see which SNPs showing strong evidence of association in the full data (that is, signals outside MHC with trend test $P < 5 \times 10^{-7}$), would have been detected at that same threshold in only a subset of our data (Fig. 6). Because it focuses on a particular but arbitrary *P*-value threshold, some care is needed in interpreting the figure. Nonetheless, for subsamples of 1,000 cases and 1,000 controls, of the 16 loci detected in the full study, we would have been certain of seeing only 2, with an expectation of about 6; for subsamples of 1,500 cases and 1,500 controls, we could expect to have seen about 9. These figures provide stark evidence that the larger the study sample, the more loci can be expected to reach threshold significance values. Indeed, given the likely distribution of effect sizes for most complex traits (see above), there are strong grounds for the prosecution of GWA studies on an even larger scale than ours, and, wherever possible, combining the results from existing GWA scans performed for the same trait. To assist such efforts, individual level data from this study will be widely available through the Consortium's Data Access Committee (follow links from http://www.wtccc.org.uk).

In our study, T1D and CD, the conditions showing strongest familial aggregation (as quantified by their sibling relative risks, λ_s), generated the largest number of highly significant associations. This relationship was not sustained in comparisons between the other five diseases. It is important to recognize that the association signals so far identified account for only a small proportion of overall familiality. There is a disparity in scale between the modest locus-specific λ_s effects attributable to the identified associations (for instance, the prominent *TCF7L2* signal for T2D translates into a λ_s of only 1.03) and the estimates of overall familiality that reflects the combined effects of all genes and shared family environment. These estimates demonstrate the limited potential of the variants thus far identified (singly or in combination) to provide clinically useful prediction of disease[117,118].

The identification and characterization of the aetiological variants that underlie replicated associations will necessitate extensive fine-mapping and functional validation. We view the WTCCC study and data set as an important first step towards harnessing the powerful molecular genomic tools now available to dissect the biological basis of common disease and translating those findings into improvements in human health.

METHODS SUMMARY

A detailed description of materials and methods is given in Methods. The workflow and organization of the project are given in Supplementary Fig. 16. Case series came from previously established collections with nationally representative recruitment: 2,000 samples were genotyped for each. The control samples came from two sources: half from the 1958 Birth Cohort and the remainder from a new UK Blood Service sample. The latter collection was established specifically for this study and is a UK national repository of anonymized DNA samples from 3,622 consenting blood donors. The vast majority of subjects were self-reported as of European Caucasian ancestry. All DNA samples were requantified and tested for degradation and PCR amplification. Genotyping was performed using GeneChip 500K arrays at the Affymetrix Services Lab (California): arrays not passing the 93% call rate threshold at $P = 0.33$ with the Dynamic Model algorithm were repeated. CEL (cell intensity) files were transferred to WTCCC for quantile normalization, and genotypes called using a new genotyping algorithm, CHIAMO, developed for this project. QC/QA measures included sample call rate, overall heterozygosity and evidence of non-European ancestry (809 samples excluded; 16,179 retained for analysis). SNPs were excluded from analysis because of missing data rates, departures from Hardy–Weinberg equilibrium and other metrics (31,011 excluded; 469,557 retained). Standard 1-d.f. and 2-d.f. tests of case-control association were supplemented with bayesian approaches, multilocus methods (data imputation) and analyses with combined data sets, either as additional cases (to detect variants influencing multiple phenotypes) or as an expanded reference group (to increase power). Results for each SNP for all analyses reported will be available from http://www.wtccc.org.uk, as will details allowing other researchers to apply for access to WTCCC genotype data. Software packages developed within the WTCCC are available on request (see Methods for details).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 26 March; accepted 11 May 2007.

1. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nature Rev. Genet.* **6,** 95–108 (2005).
2. Barrett, J. C. & Cardon, L. R. Evaluating coverage of genome-wide association studies. *Nature Genet.* **38,** 659–662 (2006).
3. The International HapMap consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
4. Murray, C. J. & Lopez, A. D. Evidence-based health policy—lessons from the Global Burden of Disease Study. *Science* **274,** 740–743 (1996).
5. Mantel, N. Chi-square tests with one degree of freedom: Extension of the Mantel–Haenszel procedure. *J. Am. Stat. Ass.* **58,** 690–700 (1963).
6. Colhoun, H. M., McKeigue, P. M. & Davey Smith, G. Problems of reporting genetic associations with complex outcomes. *Lancet* **361,** 865–872 (2003).
7. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74,** 1111–1120 (2004).
8. Coelho, M. *et al.* Microsatellite variation and evolution of human lactase persistence. *Hum. Genet.* **117,** 329–339 (2005).
9. Sabeti, P. C. *et al.* Positive natural selection in the human lineage. *Science* **312,** 1614–1620 (2006).
10. Todd, J. A. *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature Genet.* advance online publication, doi:10.1038/ng2068 (6 June 2007).
11. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164,** 1567–1587 (2003).
12. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959 (2000).
13. Menozzi, P., Piazza, A. & Cavalli-Sforza, L. Synthetic maps of human gene frequencies in Europeans. *Science* **201,** 786–792 (1978).
14. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet.* **38,** 904–909 (2006).
15. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55,** 997–1004 (1999).
16. Clayton, D. in *Handbook of Statistical Genetics* (eds Balding, D. J., Bishop, M. & Cannings, C.) 939–960 (Wiley, New York, 2003).
17. Mackay, T. F. & Anholt, R. R. Of flies and man: *Drosophila* as a model for human complex traits. *Annu. Rev. Genomics Hum. Genet.* **7,** 339–367 (2006).
18. Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L. & Rothman, N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J. Natl. Cancer Inst.* **96,** 434–442 (2004).
19. Zeggini, E. *et al.* Replication of genome-wide association signals in U.K. samples reveals risk loci for type 2 diabetes. *Science* online publication, doi:10.1126/science.1142364 (26 April 2007).
20. Steinthorsdottir, V. *et al.* A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nature Genet.* advance online publication, doi:10.1038/ng2043 (26 April 2007).
21. Scott, L. J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* online publication, doi:10.1126/science.1142382 (26 April 2007).
22. Diabetes Genetics Institute. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* online publication, doi:10.1126/science.1142358 (26 April 2007).
23. Parkes, M. *et al.* Sequence variants in the autophagy gene *IRGM* and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nature Genet.* advance online publication, doi:10.1038/ng2061 (6 June 2007).
24. Frayling, T. M. *et al.* A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316,** 889–894 (2007).
25. Cohen, J. C. *et al.* Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305,** 869–872 (2004).
26. Muller-Oerlinghausen, B., Berghofer, A. & Bauer, M. Bipolar disorder. *Lancet* **359,** 241–247 (2002).
27. Craddock, N., O'Donovan, M. C. & Owen, M. J. The genetics of schizophrenia and bipolar disorder: dissecting psychosis. *J. Med. Genet.* **42,** 193–204 (2005).
28. McGuffin, P. *et al.* The heritability of bipolar affective disorder and the genetic relationship to unipolar depression. *Arch. Gen. Psychiatry* **60,** 497–502 (2003).
29. Rice, J. *et al.* The familial transmission of bipolar illness. *Arch. Gen. Psychiatry* **44,** 441–447 (1987).
30. McQueen, M. B. *et al.* Combined analysis from eleven linkage studies of bipolar disorder provides strong evidence of susceptibility loci on chromosomes 6q and 8q. *Am. J. Hum. Genet.* **77,** 582–595 (2005).
31. Craddock, N. & Owen, M. J. The beginning of the end for the Kraepelinian dichotomy. *Br. J. Psychiatry* **186,** 364–366 (2005).
32. Ozeki, Y. *et al.* Disrupted-in-Schizophrenia-1 (*DISC-1*): mutant truncation prevents binding to NudE-like (*NUDEL*) and inhibits neurite outgrowth. *Proc. Natl Acad. Sci. USA* **100,** 289–294 (2003).

33. Blackwood, D. H. *et al.* Schizophrenia and affective disorders—cosegregation with a translocation at chromosome 1q42 that directly disrupts brain-expressed genes: clinical and P300 findings in a family. *Am. J. Hum. Genet.* **69,** 428–433 (2001).
34. Graves, T. D. & Hanna, M. G. Neurological channelopathies. *Postgrad. Med. J.* **81,** 20–32 (2005).
35. Krystal, J. H. *et al.* Glutamate and GABA systems as targets for novel antidepressant and mood-stabilizing treatments. *Mol. Psychiatry* **7** (Suppl. 1), S71–S80 (2002).
36. Vawter, M. P. *et al.* Reduction of synapsin in the hippocampus of patients with bipolar disorder and schizophrenia. *Mol. Psychiatry* **7,** 571–578 (2002).
37. Libby, P. & Theroux, P. Pathophysiology of coronary artery disease. *Circulation* **111,** 3481–3488 (2005).
38. Yusuf, S. *et al.* Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *Lancet* **364,** 937–952 (2004).
39. Lusis, A. J., Mar, R. & Pajukanta, P. Genetics of atherosclerosis. *Annu. Rev. Genomics Hum. Genet.* **5,** 189–218 (2004).
40. Watkins, H. & Farrall, M. Genetic susceptibility to coronary artery disease: from promise to progress. *Nature Rev. Genet.* **7,** 163–173 (2006).
41. Helgadottir, A. *et al.* The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke. *Nature Genet.* **36,** 233–239 (2004).
42. Helgadottir, A. *et al.* A variant of the gene encoding leukotriene A4 hydrolase confers ethnicity-specific risk of myocardial infarction. *Nature Genet.* **38,** 68–74 (2006).
43. Topol, E. J., Smith, J., Plow, E. F. & Wang, Q. K. Genetic susceptibility to myocardial infarction and coronary artery disease. *Hum. Mol. Genet.* **15** (Spec. No. 2), R117–R123 (2006).
44. Lowe, S. W. & Sherr, C. J. Tumor suppression by *Ink4a–Arf*: progress and puzzles. *Curr. Opin. Genet. Dev.* **13,** 77–83 (2003).
45. Hannon, G. J. & Beach, D. $p15^{INK4B}$ is a potential effector of TGF-β-induced cell cycle arrest. *Nature* **371,** 257–261 (1994).
46. Kalinina, N. *et al.* Smad expression in human atherosclerotic lesions: evidence for impaired TGF-β/Smad signaling in smooth muscle cells of fibrofatty lesions. *Arterioscler. Thromb. Vasc. Biol.* **24,** 1391–1396 (2004).
47. Schmid, M. *et al.* A methylthioadenosine phosphorylase (MTAP) fusion transcript identifies a new gene on chromosome 9p21 that is frequently deleted in cancer. *Oncogene* **19,** 5747–5754 (2000).
48. Prasannan, P., Pike, S., Peng, K., Shane, B. & Appling, D. R. Human mitochondrial C_1-tetrahydrofolate synthase: gene structure, tissue distribution of the mRNA, and immunolocalization in Chinese hamster ovary calls. *J. Biol. Chem.* **278,** 43178–43187 (2003).
49. Walkup, A. S. & Appling, D. R. Enzymatic characterization of human mitochondrial C_1-tetrahydrofolate synthase. *Arch. Biochem. Biophys.* **442,** 196–205 (2005).
50. Frosst, P. *et al.* A candidate genetic risk factor for vascular disease: a common mutation in methylenetetrahydrofolate reductase. *Nature Genet.* **10,** 111–113 (1995).
51. Klerk, M. *et al.* MTHFR 677C→T polymorphism and risk of coronary heart disease: a meta-analysis. *J. Am. Med. Assoc.* **288,** 2023–2031 (2002).
52. Gregory, J. F. III *et al.* Primed, constant infusion with [$^2H_3$]serine allows *in vivo* kinetic measurement of serine turnover, homocysteine remethylation, and transsulfuration processes in human one-carbon metabolism. *Am. J. Clin. Nutr.* **72,** 1535–1541 (2000).
53. Randak, C. *et al.* Three siblings with nonketotic hyperglycinaemia, mildly elevated plasma homocysteine concentrations and moderate methylmalonic aciduria. *J. Inherit. Metab. Dis.* **23,** 520–522 (2000).
54. Wight, T. N. The ADAMTS proteases, extracellular matrix, and vascular disease — Waking the sleeping giant(s)! *Arterioscler. Thromb. Vasc. Biol.* **25,** 12–14 (2005).
55. Jonsson-Rylander, A. *et al.* The role of ADAMTS-1 in atherosclerosis: Remodeling of carotid artery, immunohistochemistry, and proteolysis of versican. *Arter. Thromb. Vas. Bio.* **25,** 180–185 (2004).
56. Travis, S. P. *et al.* European evidence based consensus on the diagnosis and management of Crohn's disease: current management. *Gut* **55** (Suppl. 1), i16–i35 (2006).
57. Sartor, R. B. Mechanisms of disease: pathogenesis of Crohn's disease and ulcerative colitis. *Nature Clin. Pract. Gastroenterol. Hepatol.* **3,** 390–407 (2006).
58. Tysk, C., Lindberg, E., Jarnerot, G. & Floderusmyrhed, B. Ulcerative-colitis and Crohns-disease in an unselected population of monozygotic and dizygotic twins — a study of heritability and the influence of smoking. *Gut* **29,** 990–996 (1988).
59. Gaya, D. R., Russell, R. K., Nimmo, E. R. & Satsangi, J. New genes in inflammatory bowel disease: lessons for complex diseases? *Lancet* **367,** 1271–1284 (2006).
60. Hugot, J. P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411,** 599–603 (2001).
61. Ogura, Y. *et al.* A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411,** 603–606 (2001).
62. Rioux, J. D. *et al.* Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nature Genet.* **29,** 223–228 (2001).
63. Duerr, R. H. *et al.* A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* **314,** 1461–1463 (2006).
64. Hampe, J. *et al.* A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in *ATG16L1*. *Nature Genet.* **39,** 207–211 (2007).
65. Rioux, J. D. *et al.* Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nature Genet.* **39,** 596–604 (2007).
66. 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).
67. Singh, S. B., Davis, A. S., Taylor, G. A. & Deretic, V. Human IRGM induces autophagy to eliminate intracellular mycobacteria. *Science* **313,** 1438–1441 (2006).
68. Leonard, E. J. Biological aspects of macrophage-stimulating protein (MSP) and its receptor. *Ciba Found Symp.* **212,** 183–191; discussion 192–197 (1997).
69. Pabst, O., Forster, R., Lipp, M., Engel, H. & Arnold, H. H. NKX2.3 is required for MAdCAM-1 expression and homing of lymphocytes in spleen and mucosa-associated lymphoid tissue. *EMBO J.* **19,** 2015–2023 (2000).
70. Yamazaki, K. *et al.* Single nucleotide polymorphisms in *TNFSF15* confer susceptibility to Crohn's disease. *Hum. Mol. Genet.* **14,** 3499–3506 (2005).
71. Pietravalle, F. *et al.* Human native soluble CD40L is a biologically active trimer, processed inside microsomes. *J. Biol. Chem.* **271,** 5965–5967 (1996).
72. Battegay, E. J., Lip, G. Y. H. & Badris, G. L. (eds) *Hypertension; Principles and Practice* (Taylor Francis Group, 2005).
73. Kobberling, J. & Tattersall, R. (eds) *The Genetics of Diabetes Mellitus* (Academic Press, London, 1982).
74. Dominiczak, A. F. *et al.* Genetics of hypertension: Lessons learnt from Mendelian and polygenic syndromes. *Clin. Exp. Hypertens.* **26,** 611–620 (2004).
75. Mein, C. A., Caulfield, M. J., Dobson, R. J. & Munroe, P. B. Genetics of essential hypertension. *Hum. Mol. Genet.* **13,** R169–R175 (2004).
76. Hubner, N. *et al.* Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genet.* **37,** 243–253 (2005).
77. Caulfield, M. *et al.* Genome-wide mapping of human loci for essential hypertension. *Lancet* **361,** 2118–2123 (2003).
78. Newhouse, S. J. *et al.* Haplotypes of the *WNK1* gene associate with blood pressure variation in a severely hypertensive population from the British Genetics of Hypertension study. *Hum. Mol. Genet.* **14,** 1805–1814 (2005).
79. Tobin, M. D. *et al.* Association of *WNK1* gene polymorphisms and haplotypes with ambulatory blood pressure in the general population. *Circulation* **112,** 3423–3429 (2005).
80. Otsu, K. *et al.* Molecular cloning of cDNA encoding the Ca^{2+} release channel (ryanodine receptor) of rabbit cardiac muscle sarcoplasmic reticulum. *J. Biol. Chem.* **265,** 13472–13483 (1990).
81. Benkusky, N. A., Farrell, E. F. & Valdivia, H. H. Ryanodine receptor channelopathies. *Biochem. Biophys. Res. Commun.* **322,** 1280–1285 (2004).
82. Worthington, J., Barton, A. & John, S. L. The epidemiology of rheumatoid arthritis and the use of linkage and association studies to identify disease genes (Birkhauser, Basel, 2005).
83. Wordsworth, P. & Bell, J. Polygenic susceptibility in rheumatoid arthritis. *Ann. Rheum. Dis.* **50,** 343–346 (1991).
84. Gregersen, P. K., Silver, J. & Winchester, R. J. The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. *Arthritis Rheum.* **30,** 1205–1213 (1987).
85. Jawaheer, D. *et al.* A genomewide screen in multiplex rheumatoid arthritis families suggests genetic overlap with other autoimmune diseases. *Am. J. Hum. Genet.* **68,** 927–936 (2001).
86. John, S. *et al.* Whole-genome scan, in a complex disease, using 11,245 single-nucleotide polymorphisms: comparison with microsatellites. *Am. J. Hum. Genet.* **75,** 54–64 (2004).
87. MacKay, K. *et al.* Whole-genome linkage analysis of rheumatoid arthritis susceptibility loci in 252 affected sibling pairs in the United Kingdom. *Arthritis Rheum.* **46,** 632–639 (2002).
88. Begovich, A. B. *et al.* A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (*PTPN22*) is associated with rheumatoid arthritis. *Am. J. Hum. Genet.* **75,** 330–337 (2004).
89. Hinks, A., Eyre, S., Barton, A., Thomson, W. & Worthington, J. Investigation of genetic variation across PTPN22 in UK rheumatoid arthritis (RA) patients. *Ann. Rheum. Dis.* **66,** 683–686 (2006).
90. Sharfe, N., Dadi, H. K., Shahar, M. & Roifman, C. M. Human immune disorder arising from mutation of the α chain of the interleukin-2 receptor. *Proc. Natl Acad. Sci. USA* **94,** 3168–3171 (1997).
91. Vella, A. *et al.* Localization of a type 1 diabetes locus in the *IL2RA/CD25* region by use of tag single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **76,** 773–779 (2005).
92. Devendra, D., Liu, E. & Eisenbarth, G. S. Type 1 diabetes: recent developments. *Br. Med. J.* **328,** 750–754 (2004).
93. Hyttinen, V., Kaprio, J., Kinnunen, L., Koskenvuo, M. & Tuomilehto, J. Genetic liability of type 1 diabetes and the onset age among 22,650 young Finnish twin pairs: a nationwide follow-up study. *Diabetes* **52,** 1052–1055 (2003).
94. Smyth, D. J. *et al.* A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (*IFIH1*) region. *Nature Genet.* **38,** 617–619 (2006).
95. Todd, J. A. Statistical false positive or true disease pathway? *Nature Genet.* **38,** 731–733 (2006).
96. Mustelin, T., Vang, T. & Bottini, N. Protein tyrosine phosphatases and the immune response. *Nature Rev. Immunol.* **5,** 43–57 (2005).

97. Bottini, N. *et al.* A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. *Nature Genet.* **36,** 337–338 (2004).
98. Brand, O. J. Association of the interleukin-2 receptor alpha (*IL-2Rα*)/*CD25* gene region with Graves' disease using a multilocus test and tag SNPs. *Clin. Endocrinol.* **66,** 508–512 (2007).
99. Yamanouchi, J. *et al.* Interleukin-2 gene variation impairs regulatory T cell function and causes autoimmunity. *Nature Genet.* **39,** 329–337 (2007).
100. Zimmet, P., Alberti, K. G. & Shaw, J. Global and societal implications of the diabetes epidemic. *Nature* **414,** 782–787 (2001).
101. Stumvoll, M., Goldstein, B. J. & van Haeften, T. W. Type 2 diabetes: principles of pathogenesis and therapy. *Lancet* **365,** 1333–1346 (2005).
102. Altshuler, D. *et al.* The common PPARγ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genet.* **26,** 76–80 (2000).
103. Gloyn, A. L. *et al.* Large-scale association studies of variants in genes encoding the pancreatic β-cell KATP channel subunits Kir6.2 (*KCNJ11*) and SUR1 (*ABCC8*) confirm that the KCNJ11 E23K variant is associated with type 2 diabetes. *Diabetes* **52,** 568–572 (2003).
104. Grant, S. F. *et al.* Variant of transcription factor 7-like 2 (*TCF7L2*) gene confers risk of type 2 diabetes. *Nature Genet.* **38,** 320–323 (2006).
105. Zeggini, E. & McCarthy, M. I. *TCF7L2*: the biggest story in diabetes genetics since HLA? *Diabetologia* **50,** 1–4 (2007).
106. Helgason, A. *et al.* Refining the impact of *TCF7L2* gene variants on type 2 diabetes and adaptive evolution. *Nature Genet.* **39,** 218–225 (2007).
107. Saxena, R. *et al.* Common single nucleotide polymorphisms in *TCF7L2* are reproducibly associated with type 2 diabetes and reduce the insulin response to glucose in nondiabetic individuals. *Diabetes* **55,** 2890–2895 (2006).
108. Ubeda, M., Rukstalis, J. M. & Habener, J. F. Inhibition of cyclin-dependent kinase 5 activity protects pancreatic β cells from glucotoxicity. *J. Biol. Chem.* **281,** 28858–28864 (2006).
109. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445,** 881–885 (2007).
110. Marchini, J., Donnelly, P. & Cardon, L. R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genet.* **37,** 413–417 (2005).
111. Clayton, D. G. *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genet.* **37,** 1243–1246 (2005).
112. Zondervan, K. T. & Cardon, L. R. The complex interplay among factors that influence allelic association. *Nature Rev. Genet.* **5,** 89–100 (2004).
113. Hutchison, K. E., Stallings, M., McGeary, J. & Bryan, A. Population stratification in the candidate gene study: fatal threat or red herring? *Psychol. Bull.* **130,** 66–79 (2004).
114. Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S. & Hirschhorn, J. N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genet.* **33,** 177–182 (2003).
115. Hayes, B. & Goddard, M. E. The distribution of the effects of genes affecting quantitative traits in livestock. *Genet. Sel. Evol.* **33,** 209–229 (2001).
116. Valdar, W. *et al.* Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature Genet.* **38,** 879–887 (2006).
117. Yang, Q., Khoury, M. J., Friedman, J., Little, J. & Flanders, W. D. How many genes underlie the occurrence of common complex diseases in the population? *Int. J. Epidemiol.* **34,** 1129–1137 (2005).
118. Janssens, A. C. *et al.* Predictive testing for complex diseases using multiple genes: fact or fiction? *Genet. Med.* **8,** 395–400 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The principal funder of this project was the Wellcome Trust. Case collections were funded by: Arthritis Research Campaign, BDA Research, British Heart Foundation, British Hypertension Society, Diabetes UK, Glaxo-Smith Kline Research and Development, Juvenile Diabetes Research Foundation, National Association for Colitis and Crohn's disease, SHERT (The Scottish Hospitals Endowment Research Trust), St Bartholomew's and The Royal London Charitable Foundation, UK Medical Research Council, UK NHS R&D and the Wellcome Trust. Statistical analyses were funded by a Commonwealth Scholarship, EU, EPSRC, Fundação para a Ciência e a Tecnologia (Portugal), National Institutes of Health, National Science Foundation and the Wellcome Trust. We acknowledge the many physicians, research fellows and research nurses who contributed to the various case collections, and the collection teams and senior management of the UK Blood Services responsible for the UK Blood Services Collection. For the 1958 Birth Cohort, venous blood collection was funded by the UK Medical Research Council and cell-line production, DNA extraction and processing by the Juvenile Diabetes Research Foundation and the Wellcome Trust. We recognize the contributions of: P. Shepherd (1958 Birth Cohort); those at Affymetrix responsible for genotype assay optimization, data production and data delivery (particularly S. Cawley, R. Mei, H. Fakhrai-Rad, H. Francis-Land, R. Pillai); L. Forty, G. Fraser, J. Heron, S. Hyde, A. Massey; F. Oyebode, E. Russell, M. Sinclair, A. Stern, N. Walker and S. Zammitt (recruitment and phenotypic assessment of BD cases); M. Yuille, B. Ollier and the UK DNA Banking Network and members of the BHF Family Heart Study Research Group (CAD case recruitment and DNA provision); S. Goldthorpe, D. Soars and J. Whittaker for CD collections; J. Pembroke, M. Bruce, S. Colville-Stewart, K. Edwards, L. Gatherer, C. Gemmell, K. Gilmour, S. Hampson, S. Hood, J. Hunt, J. Hussein, J. Jamieson, J. Kent, D. Lloyd, K. MacFarlane, S. Mellow, A. Nixon, J. Pheby, D. Picton, F. Porteus, P. Whitworth, K. Witte, A. Zawadzka, C. Mein and the Barts and The London Genome Centre (HT sample collection); H. Withers, the research nurses and the membership of the British Society for Paediatric Endocrinology and Diabetes (T1D case recruitment); and M. Sampson, S. O'Rahilly, S. Howell, M. Murphy and A. Wilson (T2D case recruitment). Essential informatics support was provided by the administration, systems, bioinformatics, data services and DNA teams of the JDRF/WT DIL; the Web System teams at the Sanger Institute (particularly R. Pettitt); D. Holland and R. Vincent. T. Dibling, C. Hind, D. Simpkin, P. Ewels and D. Moore provided genotyping assistance. Personal support was provided by: Arthritis Research Campaign (A.B., H.Do., S.E., P.G., S.H., A.H., S.J., C.P., A.S., D.S., W.T., J.Wo.); British Heart Foundation (S.G.B., N.J.S., A.Do., C.W.); Cancer Research UK (D.Ea.); Diabetes UK (R.M.F.); Cure Crohn's and Colitis Fund (F.R.C.); CORE (C.M.O.); SIM (G.B.); Leverhulme Trust (A.P.M.); Throne-Holst Foundation (C.M.L.); UK Medical Research Council (D.P.K., M.D.T., J.R.P.); Vandervell Foundation (M.N.W.); and Wellcome Trust (D.G.C., L.R.C., C.M., J.Sat., M.T., A.T.H., E.Z., C.B., S.J.B., A.C., K.D., J.Gh., R.G., S.E.H., A.K., E.K., R.McG., S.P., R.R., P.Wh., D.W., P.De.).

Author Information Affymetrix GeneChip Mapping 500K Set Arrays 250K_Nsp_SNP and 250K_Sty2_SNP are deposited in NCBI GEO under accession numbers GPL3718 and GPL3720, respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.D. (donnelly@stats.ox.ac.uk).

The Wellcome Trust Case Control Consortium

Management Committee Paul R. Burton[1], David G. Clayton[2], Lon R. Cardon[3], Nick Craddock[4], Panos Deloukas[5], Audrey Duncanson[6], Dominic P. Kwiatkowski[3,5], Mark I. McCarthy[3,7], Willem H. Ouwehand[8,9], Nilesh J. Samani[10], John A. Todd[2] & Peter Donnelly (Chair)[11]

Data and Analysis Committee Jeffrey C. Barrett[3], Paul R. Burton[1], Dan Davison[11], Peter Donnelly[11], Doug Easton[12], David Evans[3], Hin-Tak Leung[2], Jonathan L. Marchini[11], Andrew P. Morris[3], Chris C. A. Spencer[11], Martin D. Tobin[1], Lon R. Cardon (Co-chair)[3] & David G. Clayton (Co-chair)[2]

UK Blood Services and University of Cambridge Controls Antony P. Attwood[5,8], James P. Boorman[8,9], Barbara Cant[8], Ursula Everson[13], Judith M. Hussey[14], Jennifer D. Jolley[8], Alexandra S. Knight[8], Kerstin Koch[8], Elizabeth Meech[15], Sarah Nutland[2], Christopher V. Prowse[16], Helen E. Stevens[2], Niall C. Taylor[8], Graham R. Walters[17], Neil M. Walker[2], Nicholas A. Watkins[8,9], Thilo Winzer[8], John A. Todd[2] & Willem H. Ouwehand[8,9]

1958 Birth Cohort Controls Richard W. Jones[18], Wendy L. McArdle[18], Susan M. Ring[18], David P. Strachan[19] & Marcus Pembrey[18,20]

Bipolar Disorder Gerome Breen[21], David St Clair[21] (Aberdeen); Sian Caesar[22], Katherine Gordon-Smith[22,23], Lisa Jones[22] (Birmingham); Christine Fraser[23], Elaine K. Green[23], Detelina Grozeva[23], Marian L. Hamshere[23], Peter A. Holmans[23], Ian R. Jones[23], George Kirov[23], Valentina Moskvina[23], Ivan Nikolov[23], Michael C. O'Donovan[23], Michael J. Owen[23], Nick Craddock[23] (Cardiff); David A. Collier[24], Amanda Elkin[24], Anne Farmer[24], Richard Williamson[24], Peter McGuffin[24] (London); Allan H. Young[25] & I. Nicol Ferrier[25] (Newcastle)

Coronary Artery Disease Stephen G. Ball[26], Anthony J. Balmforth[26], Jennifer H. Barrett[26], D. Timothy Bishop[26], Mark M. Iles[26], Azhar Maqbool[26], Nadira Yuldasheva[26], Alistair S. Hall[26] (Leeds); Peter S. Braund[10], Paul R. Burton[1], Richard J. Dixon[10], Massimo Mangino[10], Suzanne Stevens[10], Martin D. Tobin[1], John R. Thompson[1] & Nilesh J. Samani[10] (Leicester)

Crohn's Disease Francesca Bredin[27], Mark Tremelling[27], Miles Parkes[27] (Cambridge); Hazel Drummond[28], Charles W. Lees[28], Elaine R. Nimmo[28], Jack Satsangi[28] (Edinburgh); Sheila A. Fisher[29], Alastair Forbes[30], Cathryn M. Lewis[29], Clive M. Onnie[29], Natalie J. Prescott[29], Jeremy Sanderson[31], Christopher G. Mathew[29] (London); Jamie Barbour[32], M. Khalid Mohiuddin[32], Catherine E. Todhunter[32], John C. Mansfield[32] (Newcastle); Tariq Ahmad[33], Fraser R. Cummings[33] & Derek P. Jewell[33] (Oxford)

Hypertension John Webster[34] (Aberdeen); Morris J. Brown[35], David G. Clayton[2] (Cambridge); G. Mark Lathrop[36] (Evry); John Connell[37], Anna Dominiczak[37] (Glasgow); Nilesh J. Samani[10] (Leicester); Carolina A. Braga Marcano[38], Beverley Burke[38], Richard Dobson[38], Johannie Gungadoo[38], Kate L. Lee[38], Patricia B. Munroe[38], Stephen J. Newhouse[38], Abiodun Onipinla[38], Chris Wallace[38], Mingzhan Xue[38], Mark Caulfield[38] (London); Martin Farrall[39] (Oxford)

Rheumatoid Arthritis Anne Barton[40], The Biologics in RA Genetics and Genomics Study Syndicate (BRAGGS) Steering Committee*, Ian N. Bruce[40], Hannah Donovan[40], Steve Eyre[40], Paul D. Gilbert[40], Samantha L. Hider[40], Anne M. Hinks[40], Sally L. John[40],

Catherine Potter[40], Alan J. Silman[40], Deborah P. M. Symmons[40], Wendy Thomson[40] & Jane Worthington[40]

Type 1 Diabetes David G. Clayton[2], David B. Dunger[2,41], Sarah Nutland[2], Helen E. Stevens[2], Neil M. Walker[2], Barry Widmer[2,41] & John A. Todd[2]

Type 2 Diabetes Timothy M. Frayling[42,43], Rachel M. Freathy[42,43], Hana Lango[42,43], John R. B. Perry[42,43], Beverley M. Shields[43], Michael N. Weedon[42,43], Andrew T. Hattersley[42,43] (Exeter); Graham A. Hitman[44] (London); Mark Walker[45] (Newcastle); Kate S. Elliott[3,7], Christopher J. Groves[7], Cecilia M. Lindgren[3,7], Nigel W. Rayner[3,7], Nicholas J. Timpson[3,46], Eleftheria Zeggini[3,7] & Mark I. McCarthy[3,7] (Oxford)

Tuberculosis Melanie Newport[47], Giorgio Sirugo[47] (Gambia); Emily Lyons[3], Fredrik Vannberg[3] & Adrian V. S. Hill[3] (Oxford)

Ankylosing Spondylitis Linda A. Bradbury[48], Claire Farrar[49], Jennifer J. Pointon[48], Paul Wordsworth[49] & Matthew A. Brown[48,49]

Autoimmune Thyroid Disease Jayne A. Franklyn[50], Joanne M. Heward[50], Matthew J. Simmonds[50] & Stephen C. L. Gough[50]

Breast Cancer Sheila Seal[51], Breast Cancer Susceptibility Collaboration (UK)*, Michael R. Stratton[51,52] & Nazneen Rahman[51]

Multiple Sclerosis Maria Ban[53], An Goris[53], Stephen J. Sawcer[53] & Alastair Compston[53]

Gambian Controls David Conway[47], Muminatou Jallow[47], Melanie Newport[47], Giorgio Sirugo[47] (Gambia); Kirk A. Rockett[3] & Dominic P. Kwiatkowski[3,5] (Oxford)

DNA, Genotyping, Data QC and Informatics Suzannah J. Bumpstead[5], Amy Chaney[5], Kate Downes[2,5], Mohammed J. R. Ghori[5], Rhian Gwilliam[5], Sarah E. Hunt[5], Michael Inouye[5], Andrew Keniry[5], Emma King[5], Ralph McGinnis[5], Simon Potter[5], Rathi Ravindrarajah[5], Pamela Whittaker[5], Claire Widden[5], David Withers[5], Panos Deloukas[5] (Wellcome Trust Sanger Institute, Hinxton); Hin-Tak Leung[2], Sarah Nutland[2], Helen E. Stevens[2], Neil M. Walker[2] & John A. Todd[2] (Cambridge)

Statistics Doug Easton[12], David G. Clayton[2] (Cambridge); Paul R. Burton[1], Martin D. Tobin[1] (Leicester); Jeffrey C. Barrett[3], David Evans[3], Andrew P. Morris[3], Lon R. Cardon[3] (Oxford) Niall J. Cardin[11], Dan Davison[11], Teresa Ferreira[11], Joanne Pereira-Gale[11], Ingileif B. Hallgrimsdóttir[11], Bryan N. Howie[11], Jonathan L. Marchini[11], Chris C. A. Spencer[11], Zhan Su[11], Yik Ying Teo[3,11], Damjan Vukcevic[11] & Peter Donnelly[11] (Oxford)

Primary Investigators David Bentley[5]†, Matthew A. Brown[48,49], Lon R. Cardon[3], Mark Caulfield[38], David G. Clayton[2], Alistair Compston[53], Nick Craddock[23], Panos Deloukas[5], Peter Donnelly[11], Martin Farrall[39], Stephen C. L. Gough[50], Alistair S. Hall[26], Andrew T. Hattersley[42,43], Adrian V. S. Hill[3], Dominic P. Kwiatkowski[3,5], Christopher G. Mathew[29], Mark I. McCarthy[3,7], Willem H. Ouwehand[8,9], Miles Parkes[27], Marcus Pembrey[18,20], Nazneen Rahman[51], Nilesh J. Samani[10], Michael R. Stratton[51,52], John A. Todd[2] & Jane Worthington[40]

*See Supplementary Information for details.

Affiliations for participants: [1]Genetic Epidemiology Group, Department of Health Sciences, University of Leicester, Adrian Building, University Road, Leicester LE1 7RH, UK. [2]Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Wellcome Trust/MRC Building, Cambridge CB2 0XY, UK. [3]Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. [4]Department of Psychological Medicine, Henry Wellcome Building, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK. [5]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [6]The Wellcome Trust, Gibbs Building, 215 Euston Road, London NW1 2BE, UK. [7]Oxford Centre for Diabetes, Endocrinology and Medicine, University of Oxford, Churchill Hospital, Oxford OX3 7LJ, UK. [8]Department of Haematology, University of Cambridge, Long Road, Cambridge CB2 2PT, UK. [9]National Health Service Blood and Transplant, Cambridge Centre, Long Road, Cambridge CB2 2PT, UK. [10]Department of Cardiovascular Sciences, University of Leicester, Glenfield Hospital, Groby Road, Leicester LE3 9QP, UK. [11]Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK. [12]Cancer Research UK Genetic Epidemiology Unit, Strangeways Research Laboratory, Worts Causeway, Cambridge CB1 8RN, UK. [13]National Health Service Blood and Transplant, Sheffield Centre, Longley Lane, Sheffield S5 7JN, UK. [14]National Health Service Blood and Transplant, Brentwood Centre, Crescent Drive, Brentwood CM15 8DP, UK. [15]The Welsh Blood Service, Ely Valley Road, Talbot Green, Pontyclun CF72 9WB, UK. [16]The Scottish National Blood Transfusion Service, Ellen's Glen Road, Edinburgh EH17 7QT, UK. [17]National Health Service Blood and Transplant, Southampton Centre, Coxford Road, Southampton SO16 5AF, UK. [18]Avon Longitudinal Study of Parents and Children, University of Bristol, 24 Tyndall Avenue, Bristol BS8 1TQ, UK. [19]Division of Community Health Services, St George's University of London, Cranmer Terrace, London SW17 0RE, UK. [20]Institute of Child Health, University College London, 30 Guilford Street, London WC1N 1EH, UK. [21]University of Aberdeen, Institute of Medical Sciences, Foresterhill, Aberdeen AB25 2ZD, UK. [22]Department of Psychiatry, Division of Neuroscience, Birmingham University, Birmingham B15 2QZ, UK. [23]Department of Psychological Medicine, Henry Wellcome Building, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK. [24]SGDP, The Institute of Psychiatry, King's College London, De Crespigny Park, Denmark Hill, London SE5 8AF, UK. [25]School of Neurology, Neurobiology and Psychiatry, Royal Victoria Infirmary, Queen Victoria Road, Newcastle upon Tyne, NE1 4LP, UK. [26]LIGHT and LIMM Research Institutes, Faculty of Medicine and Health, University of Leeds, Leeds LS1 3EX, UK. [27]IBD Research Group, Addenbrooke's Hospital, University of Cambridge, Cambridge CB2 2QQ, UK. [28]Gastrointestinal Unit, School of Molecular and Clinical Medicine, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK. [29]Department of Medical & Molecular Genetics, King's College London School of Medicine, 8th Floor Guy's Tower, Guy's Hospital, London SE1 9RT, UK. [30]Institute for Digestive Diseases, University College London Hospitals Trust, London, NW1 2BU, UK. [31]Department of Gastroenterology, Guy's and St Thomas' NHS Foundation Trust, London SE1 7EH, UK. [32]Department of Gastroenterology & Hepatology, University of Newcastle upon Tyne, Royal Victoria Infirmary, Newcastle upon Tyne NE1 4LP, UK. [33]Gastroenterology Unit, Radcliffe Infirmary, University of Oxford, Oxford OX2 6HE, UK. [34]Medicine and Therapeutics, Aberdeen Royal Infirmary, Foresterhill, Aberdeen, Grampian AB9 2ZB, UK. [35]Clinical Pharmacology Unit and the Diabetes and Inflammation Laboratory, University of Cambridge, Addenbrookes Hospital, Hills Road, Cambridge CB2 2QQ, UK. [36]Centre National de Genotypage, 2, Rue Gaston Cremieux, Evry, Paris 91057, France. [37]BHF Glasgow Cardiovascular Research Centre, University of Glasgow, 126 University Place, Glasgow G12 8TA, UK. [38]Clinical Pharmacology and Barts and The London Genome Centre, William Harvey Research Institute, Barts and The London, Queen Mary's School of Medicine, Charterhouse Square, London EC1M 6BQ, UK. [39]Cardiovascular Medicine, University of Oxford, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK. [40]arc Epidemiology Research Unit, University of Manchester, Stopford Building, Oxford Rd, Manchester M13 9PT, UK. [41]Department of Paediatrics, University of Cambridge, Addenbrooke's Hospital, Cambridge CB2 2QQ, UK. [42]Genetics of Complex Traits, Institute of Biomedical and Clinical Science, Peninsula Medical School, Magdalen Road, Exeter EX1 2LU, UK. [43]Diabetes Genetics, Institute of Biomedical and Clinical Science, Peninsula Medical School, Barrack Road, Exeter EX2 5DU, UK. [44]Centre for Diabetes and Metabolic Medicine, Barts and The London, Royal London Hospital, Whitechapel, London E1 1BB, UK. [45]Diabetes Research Group, School of Clinical Medical Sciences, Newcastle University, Framlington Place, Newcastle upon Tyne NE2 4HH, UK. [46]The MRC Centre for Causal Analyses in Translational Epidemiology, Bristol University, Canynge Hall, Whiteladies Rd, Bristol BS2 8PR, UK. [47]MRC Laboratories, Fajara, The Gambia. [48]Diamantina Institute for Cancer, Immunology and Metabolic Medicine, Princess Alexandra Hospital, University of Queensland, Woolloongabba, Qld 4102, Australia. [49]Botnar Research Centre, University of Oxford, Headington, Oxford OX3 7BN, UK. [50]Department of Medicine, Division of Medical Sciences, Institute of Biomedical Research, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK. [51]Section of Cancer Genetics, Institute of Cancer Research, 15 Cotswold Road, Sutton SM2 5NG, UK. [52]Cancer Genome Project, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [53]Department of Clinical Neurosciences, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge CB2 2QQ, UK. †Present address: Illumina Cambridge, Chesterford Research Park, Little Chesterford, Nr Saffron Walden, Essex CB10 1XL, UK.

METHODS

BD phenotype description. BD cases were all over the age of 16 yr, living in mainland UK and of European descent. Recruitment was undertaken throughout the UK by teams based in Aberdeen (8% of cases), Birmingham (35% cases), Cardiff (33% cases), London (15% cases) and Newcastle (9% cases). Individuals who had been in contact with mental health services were recruited if they suffered with a major mood disorder in which clinically significant episodes of elevated mood had occurred. This was defined as a lifetime diagnosis of a bipolar mood disorder according to Research Diagnostic Criteria[119] and included the bipolar subtypes that have been shown in family studies to co-aggregate for example[29]: bipolar I disorder (71% cases), schizoaffective disorder bipolar type (15% cases), bipolar II disorder (9% cases) and manic disorder (5% cases). After providing written informed consent, all subjects were interviewed by a trained psychologist or psychiatrist using a semi-structured lifetime diagnostic psychiatric interview (in most cases the Schedules for Clinical Assessment in Neuropsychiatry[120] and available psychiatric medical records were reviewed). Using all available data, best-estimate ratings were made for a set of key phenotypic measures on the basis of the OPCRIT checklist (which covers both psychopathology and course of illness)[121,122] and lifetime psychiatric diagnoses were assigned according to the Research Diagnostic Criteria[119]. The reliability of these methods has been shown to be high[119,123,124]. Further details of clinical methodology can be found in Green, 2005 (ref. 123) and Green, 2006 (ref. 124).

CAD phenotype description. CAD cases had a validated history of either myocardial infarction or coronary revascularization (coronary artery bypass surgery or percutaneous coronary angioplasty) before their 66th birthday. Verification of the history of CAD was required either from hospital records or the primary care physician. Recruitment was carried out on a national basis in the UK through a direct approach to the public via (1) the media and (2) mailing all general practices (family physicians) with information about the study, as previously described[125]. In an initial pilot phase, potential participants were also identified and approached through local CAD databases in the two lead centres (Leeds and Leicester). Although the majority of subjects had at least one further sib also affected with premature CAD, only one subject from each family was included in the present study.

CD phenotype description. CD cases were attendees at inflammatory bowel disease clinics in and around the five centres which contributed samples to the WTCCC (Cambridge, Oxford, London, Newcastle, Edinburgh). Ascertainment was based on a confirmed diagnosis of Crohn's disease (CD) using conventional endoscopic, radiological and histopathological criteria[126]. We included all subtypes of CD as classified by disease extent and behaviour and the collection was not specifically enriched for family history or early age of onset. The median age of diagnosis was 26.1 yr and 62% of the collection had undergone CD-related abdominal surgery. A small proportion had previously been recruited as members of multiply affected families but only one affected individual was included per family.

HT phenotype description. HT cases comprised severely hypertensive probands ascertained from families with multiplex affected sibships or as parent–offspring trios. They were of white British ancestry (up to level of grand-parents) and were recruited from the Medical Research Council General Practice Framework and other primary care practices in the UK[77]. Each case had a history of hypertension diagnosed before 60 yr of age, with confirmed blood pressure recordings corresponding to seated levels $>150/100$ mm Hg (if based on one reading), or the mean of 3 readings greater than 145/95 mm Hg. These criteria correspond to the threshold for the uppermost 5% of blood pressure distribution in a contemporaneous health screening survey of 5,000 British men and women in 1995 (N. Wald and M. Law, personal communication). We excluded hypertensive individuals who self-reportedly consumed >21 units of alcohol per week and those with diabetes, intrinsic renal disease, a history of secondary hypertension or co-existing illness. Cases did not undergo systematic genetic screening to exclude the (rare) known monogenic causes of HT. We focused on the recruitment of hypertensive individuals with body mass indices $<30\,kg\,m^{-2}$. The probands were extensively phenotyped by trained nurses (see http://www.brightstudy.ac.uk for standard operating procedures, additional phenotypes and study questionnaires). Sample selection for WTCCC was based on DNA availability and quality.

RA phenotype description. RA cases were recruited to studies coordinated by the ARC (Arthritis Research Campaign) Epidemiology Unit. All subjects were Caucasian over the age of 18 yr and satisfied the 1987 American College of Rheumatology Criteria for RA[127] modified for genetic studies[128]. Of the cases, 404 were recruited as part of the arc National Repository of Family Material[129]: of these, 301 were probands from affected sibling pair families and 103 were cases from trio families, having both parents or one parent and one unaffected sibling available for study. A further 109 cases were recruited from the Norfolk Arthritis Register, a primary care-based inception collection[130]. All other cases ($n = 1348$) were recruited from NHS Rheumatology Clinics throughout the UK. Samples for WTCCC were selected from the various studies on the basis of the quality and availability of DNA.

T1D phenotype description. T1D cases were recruited from paediatric and adult diabetes clinics at 150 National Health Service hospitals across mainland UK. The total T1D case data set ($n=\sim$8,000) from which the WTCCC cases were selected, represents close to half the T1D cases seen in such clinics. Nationwide coverage was achieved through the voluntary efforts of members of the British Society for Paediatric Endocrinology and Diabetes, who recruited about half of cases, the rest coming from peripatetic nurses employed by the JDRF/WT GRID project (http://www-gene.cimr.cam.ac.uk/todd/)[131]. To establish a positive diagnosis of T1D (and, in particular, to distinguish it from the more common, but later onset T2D), we required all cases to have an age of diagnosis below 17 yr and insulin dependence since diagnosis (with a minimum period of at least 6 months). However, a very few subjects were subsequently discovered to be suffering from rare monogenic disorders, such as maturity onset diabetes of the young (MODY), and latterly permanent neonatal diabetes (PNDM): these were excluded.

T2D phenotype description. The T2D cases were selected from UK Caucasian subjects who form part of the Diabetes UK Warren 2 repository. In each case, the diagnosis of diabetes was based on either current prescribed treatment with sulphonylureas, biguanides, other oral agents and/or insulin or, in the case of individuals treated with diet alone, historical or contemporary laboratory evidence of hyperglycaemia (as defined by the World Health Organization). Other forms of diabetes (for example, maturity-onset diabetes of the young, mitochondrial diabetes, and type 1 diabetes) were excluded by standard clinical criteria based on personal and family history. Criteria for excluding autoimmune diabetes included absence of first-degree relatives with T1D, an interval of $\geq$1 yr between diagnosis and institution of regular insulin therapy and negative testing for antibodies to glutamic acid decarboxylase (anti-GAD). Cases were limited to those who reported that all four grandparents had exclusively British and/or Irish origin, by both self-reported ethnicity and place of birth. All were diagnosed between age 25 and 75. Approximately 30% were explicitly recruited as part of multiplex sibships[132] and $\sim$25% were offspring in parent–offspring 'trios' or 'duos' (that is, families comprising only one parent complemented by additional sibs)[133]. The remainder were recruited as isolated cases but these cases were (compared to population-based cases) of relatively early onset and had a high proportion of T2D parents and/or siblings[134]. Cases were ascertained across the UK but were centred around the main collection centres (Exeter, London, Newcastle, Norwich, Oxford). Selection of the samples typed in WTCCC from the larger collections was based primarily on DNA availability and success in passing Diabetes and Inflammation Laboratory (DIL)/Wellcome Trust Sanger Institute (WTSI) DNA quality control.

1958 Birth Cohort Controls (58BC). The 1958 Birth Cohort (also known as the National Child Development Study) includes all births in England, Wales and Scotland, during one week in 1958. From an original sample of over 17,000 births, survivors were followed up at ages 7, 11, 16, 23, 33 and 42 yr (http://www.cls.ioe.ac.uk/studies.asp?section=000100020003)[135]. In a biomedical examination at 44-45 yr[136] (http://www.b58cgene.sgul.ac.uk/followup.php), 9,377 cohort members were visited at home providing 7,692 blood samples with consent for future Epstein–Barr virus (EBV)-transformed cell lines. DNA samples extracted from 1,500 cell lines of self-reported white ethnicity and representative of gender and each geographical region were selected for use as controls.

UK Blood Services Controls (UKBS). The second set of common controls was made up of 1,500 individuals selected from a sample of blood donors recruited as part of the current project. WTCCC in collaboration with the UK Blood Services (NHSBT in England, SNBTS in Scotland and WBS in Wales) set up a UK national repository of anonymized samples of DNA and viable mononuclear cells from 3,622 consenting blood donors, age range 18–69 yr (ethical approval 05/Q0106/74). A set of 1,564 samples was selected from the 3622 samples recruited based on sex and geographical region (to reproduce the distribution of the samples of the 1958 Birth Cohort) for use as common controls in the WTCCC study. DNA was extracted as described below with a yield of $3054 \pm 1207\,\mu g$ (mean $\pm$ 1 s.d.).

Protocol for DNA extraction. White blood cells were isolated from the filters by first pushing 10 ml air through the filter in contra direction to the initial blood flow through the filter, followed by 40 ml PBS, collecting into a 50 ml centrifuge tube, and centrifugation (2.000 r.p.m., 10 min, 20 °C).Cells were lysed by adding 40 ml Lysis buffer (320 mM Sucrose, 1% Triton-X-100, 4.9 mM $MgCl_2$, 1 mM TRIS-HCl pH 7.4) and pelleted by centrifugation (2,500 r.p.m., 15 min, 4 °C). Pellets were frozen before extraction. Pellets were digested overnight at 37 °C with 5.25 M GuHCl, 490 mM NH_4Ac, 1.25% Na Sarcosyl and $0.125\,mg\,ml^{-1}$ Proteinase K and then mixed with 2 ml chloroform to form a white emulsion. The aqueous layer was separated by centrifugation (2,500 r.p.m., 3 min) and

DNA was precipitated in ethanol overnight at −20 °C. DNA was further precipitated by rotation (40 r.p.m., 5 min) and then pelleted by centrifugation (3,000 r.p.m., 15 min). Pellets were washed twice by rinsing with 2 ml 70% ethanol, followed by centrifugation (3,000 r.p.m., 5 min). DNA pellets were air-dried before re-suspension in TE buffer (10 mM Tris, 0.1 mM EDTA).

Sample handling. Each participating sample collection was issued unique WTCCC barcode labels and a spreadsheet with unique sample identifiers for logging information on case/control status, DNA concentration (requested at 100 ng μl^{-1}), DNA extraction method, sex, broad geographical region and age at requirement. Each collection supplied 10 μg aliquots of anonymized samples in bar-coded, deep 96-well plates. On receipt, samples had their DNA concentration measured by Picogreen (triplicate measurements), were checked for DNA degradation on a 0.75% agarose gel, and genotyped with up to 38 SNPs arranged in two multiplex reactions using the MassExtend (hME) and/or iPLEX[37] assay. The above SNPs served for obtaining a molecular fingerprint (25 of the 38 SNPs were present on the GeneChip 500K) and experimentally confirming the sex of each sample.

Samples with concentrations ≥50 ng μl^{-1}, showing limited or no degradation, having a minimum of 7/10 (hME reaction) and/or 14/23 (iPLEX reaction) SNPs typed, and having the sex markers in agreement or not violating the supplied information were deemed fit for whole genome genotyping. Note that the hME set was replaced with a second iPLEX reaction in the course of the project to increase marker density. We selected 2,000 and 1,500 samples from each disease and control collection respectively. Selected samples were normalized to 50 ng μl^{-1} and re-arrayed robotically into 96-well plates so that each plate was composed of 94 samples representing at least two different collections at a ratio of 1:1. For each collection, the selected samples were balanced first for sex and then geographical region (see above).

Genotyping. SNP genotyping was performed with the commercial release of the GeneChip 500K arrays at Affymetrix Services Lab. A modified version of the genotyping assay developed for the 100K Mapping Array[137] was used. In brief, two aliquots of 250 ng of DNA each are digested with *Nsp*I and *Sty*I, respectively, an adaptor is ligated and molecules are then fragmented and labelled. At this stage each enzyme preparation is hybridized to the corresponding SNP array (262,000 and 238,000 on the *Nsp*I and *Sty*I array respectively). Samples were processed in 96-well plate format, each plate carried a positive and a negative control, up to the hybridization step. Individual arrays not passing the 93% call rate threshold at $P = 0.33$ with the Dynamic Model algorithm[138] were repeated (fresh aliquot of initial end-labelled reaction). Samples failing twice at the hybridization stage were reprocessed using a fresh DNA aliquot. Affymetrix delivered successful samples as those having a Dynamic Model call rate of 93% at $P = 0.33$ for each array, over 90% concordance for the 50 SNPs that are common to the two arrays, both arrays agreed on gender, and showed over 70% identity to the Sequenom genotypes supplied by WTCCC.

CEL files provided the intensities of the various probes on each chip. Initially, genotypes were called with the Dynamic Model[138] algorithm. Affymetrix subsequently developed an improved algorithm, BRLMM (Bayesian Robust Linear Model with Mahalanobis distance classifier[139,140]). This processes batches of samples and uses clustering techniques to call genotypes (the 'mismatch' probe intensities are not used). In Affymetrix's standard protocol it is applied in batches of 96 samples (plates). This is, of course, a very small sample size and, for some SNPs, some clusters will contain few, if any, observations. This might be countered by combining information about cluster location over a large number of SNPs.

Throughout, physical coordinates refer to NCBI build-35 of the human genome. Alleles are expressed in the forward (+) strand of the reference human genome (NCBI build-35).

Power calculations. We assessed power of the Affymetrix 500K chip using the following simulation experiment. Separately for each SNP with MAF >5% in the 10 HapMap ENCODE regions, we assumed the SNP was causative and simulated genotype data at all SNPs in the same region as the putative disease SNP in case-control panels of 2,000 cases and 3,000 controls with linkage disequilibrium patterns that match those in HapMap. For controls, these simulations were based on the imputation algorithm described below (with all genotype data initially set to missing in the 3,000 control individuals). For cases, the assumed effect size was first used to calculate genotype frequencies in cases (via Bayes' theorem), and genotypes in cases at the putative SNP were then simulated independently from theses calculated frequencies. Genotypes at all other SNPs in the region in cases were then simulated using the imputation algorithm described below (with all data other than the genotypes at the causative SNP initially set to missing in the cases). For each such simulated case-control panel, trend tests were performed at each of the SNPs in the region that are actually on the Affymetrix chip, and if any of these reached the stated *P*-value threshold the putative disease SNP was deemed to be detected, and otherwise to be undetected. Power estimates are then calculated as the proportion of putative disease SNPs with MAFs >5% across the HapMap ENCODE regions that are detected at the given *P*-value threshold. There are various approximations here. Actual numbers of cases and controls for each disease are slightly smaller than the 3,000:2,000 values used in the simulations, but in the other direction, our simulations ignore the possibility that a disease SNP might be detected by a genotyped SNP outside its ENCODE region. The accuracy reported below of the imputation algorithm in imputing genotypes leads us to believe these simulations should be a reasonable proxy for real data. Some such simulation is needed if power calculations are to take account of the fact that any given putative disease SNP could typically be detected by several SNPs on the chip. Exploitation of this simulation approach to assess power across different platforms and SNP chips and for different experimental designs will be reported elsewhere.

CHIAMO. We developed a new genotype calling algorithm, CHIAMO, which is applied after quantile normalization of the data from each sample. A complete description is given in Supplementary Information. We briefly summarize some features here. Normalized intensities for each genotype were mapped to a two-dimensional intensity vector and then we applied CHIAMO, which uses a bayesian hierarchical 4-class mixture model to call genotypes for the whole project. We used optimization based on 12 random starts to find the set of parameters ($\hat{\theta}$) that maximize the posterior distribution of the model. This parameter set was used to calculate the maximum a posteriori estimates of the probabilities of each genotype call, $\Pr\left(Z_{ij} \middle| Data, \hat{\theta}\right)$, where $Z_{ij} \in \{0, 1, 2, 3\} \equiv \{AA, AB, BB, null\}$ is the genotype call for individual *j* in collection *i*. All CHIAMO genotype calls analysed in this paper were based on an a posteriori probability threshold of 0.9 for making a call, following our analysis of the relationship between concordance and missing data rates (data not shown). CHIAMO differs from BRLMM in several respects: (1) it uses a different transformation of the CEL files to give the two-dimensional summary for each individual at an SNP leading to better defined clusters; (2) it makes use of mis-match probe signals; (3) it uses a different method for fitting the clusters; and (4) it allows the data for all samples to be called simultaneously, thus allowing better estimation of cluster location and shape parameters, while making allowance for possible differences in these parameter values between case/control groups that could arise as a result of differences in DNA quality. This is achieved using a hierarchical statistical model that specifies the joint distribution of the three cluster centres, their spread, and likely allele frequencies (using HapMap) and genotype frequencies (centred on Hardy–Weinberg proportions but allowing some variation).

CHIAMO improved both call rate and accuracy in comparison to BRLMM, the current standard Affymetrix calling algorithm (Supplementary Table 3)—it roughly halved missing data rates and discordance rates with another platform. See Supplementary Information for full details, discussion of some challenges for genotype calling, and example cluster plots (Supplementary Figs 10 and 17).

Quantile-quantile plots. Quantile-quantile (Q-Q) plots are constructed by ranking a set of values of a statistic from smallest to largest (the 'order statistics') and plotting them against their expected values, given the assumption that the values have been sampled from a distribution of known theoretical form (in our case, the chi-squared distribution, usually on one degree of freedom—for example, the distribution of our trend tests under the null hypothesis). Deviations from the line of equality indicate either that the theoretical distribution is incorrect, or that the sample is contaminated with values generated in some other manner (for example, by a true association). To aid interpretation of such plots we have also calculated 95% 'concentration bands' (shaded grey in all Q-Q plots). These are formed by calculating, for each order statistic, the 2.5th and 97.5th centiles of the distribution of the order statistic under random sampling and the null hypothesis (for details see ref. 141). We should add two notes of caution. First, concentration bands are calculated point by point and, although there are very strong correlations between nearby order statistics, the probability that a real quantile-quantile plot will stray outside the concentration band at some point is some bit larger than 5%. Second, the theoretical chi-squared distribution is an approximation, valid for large samples; it is not clear whether this approximation continues to hold into the extreme right hand tail of the distribution explored in a GWA study (although the indications are that it is probably not far wrong for a study as large as ours).

Data quality control. Of samples for which Affymetrix returned CEL files, a total of 809 were excluded from the analysis. A complete breakdown by collection is given in Supplementary Table 4. Missing data rate per sample acts as an indicator of low DNA quality. Most samples had very low rates of missing data (study-wide average 0.00925, standard deviation 0.0187) and we chose to exclude 250 samples with >3% missing data across all SNPs (Supplementary Fig. 18, and Supplementary Tables 4 and 13). We also set empirical thresholds on genome-wide heterozygosity (excess heterozygosity in particular may indicate contamination). Six samples with >30% heterozygosity and a further three with <23% heterozygosity were excluded (see Supplementary Fig. 18). We excluded 16 samples

with discrepancies between WTCCC information and external identifying information (such as genotypes from another experiment, blood type or incorrect disease status). We sought to detect individuals with non-Caucasian ancestry using multi-dimensional scaling to provide a two-dimensional projection of the data whose axes represent geographic genetic variation. In the interest of computational efficiency and to avoid confounding of the multi-dimensional scaling by extended linkage disequilibrium we thinned the data to a set of 71,458 SNPs, within which no pair were correlated with $r^2 > 0.2$. For this set of nearly independent SNPs we computed genome-wide average identity by state (sum of the number of identical-by-state alleles at each locus divided by twice the number of loci) between each pair of individuals in each sample collection along with the 270 HapMap samples. We converted these identity by-state-relationships to distances by subtracting them from 1, and the matrix of pairwise identity by state values was used as input to multi-dimensional scaling. The projection onto the two multi-dimensional scaling axes is shown in Supplementary Fig. 5. We excluded 153 samples that were clearly separate from the main cluster of WTCCC individuals. Exclusion of these individuals resulted in a substantial reduction in estimates of over-dispersion in test statistic distributions (data not shown). We also excluded 295 duplicated (>99% identity) and 86 related (86–98% identity) samples from the analysis.

Filtering out suboptimal markers depends on both the platform and the genotype calling algorithm. We experimented with various quality metrics for CHIAMO calls, for example, based on the location and/or separation of the clusters, but found that the best indicator of a SNP being difficult to call was the amount of missing data in its calls: CHIAMO consistently marked many individuals missing for SNPs with poorly defined or overlapping clusters, whereas it successfully called genotypes for nearly all individuals on high-quality SNPs (data not shown). We excluded 26,567 SNPs with a study-wide missing data rate >5% (Supplementary Fig. 19), or >1% for SNPs with a study-wide MAF < 5%. We additionally excluded 4,351 SNPs with Hardy–Weinberg exact P value $< 5.7 \times 10^{-7}$ in the combined set of 2,938 controls, and 93 SNPs with P value $< 5.7 \times 10^{-7}$ for either a one- or two-degree of freedom test of association between the two control groups (corresponding to a 1 d.f. chi-squared statistic of about 25). See Supplementary Fig. 20 and Fig. 1 respectively for the empirical distributions of these statistics used to motivate the thresholds above.

Overall, we found that the 809 excluded individuals (which represent 4.8% of the study samples) accounted for 35.6% of the missing data at non-excluded SNPs. In total, 469,557 SNPs passed the quality control filters.

Supplementary Fig. 20 shows the effect of quality control filters, and visual inspection of the cluster plots of SNPs showing apparently strong association, on quantile-quantile plots for one disease (T2D, others are similar), and the success of these filters in excluding poorly performing SNPs. The figure (panel d) also shows the marked effect on the tails of the distribution of test statistics of regions of genuine association (for this disease the three regions removed because of strong evidence of association have all been independently replicated, see main text). The aim in filtering is to exclude poor SNPs but without removing genuine associations. No single criterion will do this. In order not to exclude possible genuine associations, we chose to apply relatively light quality control filters but then to subject all apparently associated SNPs to visual inspection of cluster plots (see Supplementary Information). Around 100 cluster plots were assessed per disease.

We used X-chromosome SNPs to check for sex discrepancies with the sample files (Supplementary Fig. 21). These were fed back to disease groups for amendment and verification. The ~80 samples where it was not possible to discern the source of the discrepancy were left in the study for analysis, on the grounds that mishandling was considered unlikely to have introduced samples with altogether different phenotypes.

DNA quality between cases and controls could result in false-positive associations through differential effects on genotype calling[111]. DNAs in our study came from various sources between, and in some cases within, case and control series, but with the combination of centralized sample quality control, simultaneous genotype calling with CHIAMO (which explicitly allows for differences between collections), and inspection of cluster plots for SNPs with very small P values, our study did not experience such difficulties.

Comparing linkage disequilibrium. Two questions which have been raised about the HapMap data are how well it describes linkage disequilibrium in populations other than the ones that were sampled, and whether the sample sizes in HapMap (60 Caucasian individuals, for example) are adequate to describe patterns of linkage disequilibrium. With data on 2,938 controls and 16,179 individuals in total at around 400,000 polymorphic SNPs, we are well placed to address this for the British population. Initial analyses suggest that patterns of linkage disequilibrium in our samples are very similar to those in HapMap. As an example, Supplementary Fig. 3 compares patterns of linkage disequilibrium in HapMap CEU individuals and our 58C sample at SNPs on the Affymetrix chip across 22× 1 Mb regions of the genome and they seem almost identical. We calculated r^2 values directly from the phased haplotypes available in HapMap, but using unphased genotype data from our study. Note that visual representations of linkage disequilibrium in this form can be very sensitive to SNP density so comparisons across regions is difficult without correction for SNP density, and direct comparison of linkage disequilibrium patterns at all HapMap SNPs with those at the subset of SNPs on the Affymetrix 500K chip is not straightforward.

Geographical variation and population structure. Principal component analysis was performed as a two-stage process: we formed a matrix of estimated correlations (formally, the inner product measure of similarity) between all pairs of individuals, and then computed the eigenvectors and eigenvalues of that matrix. We estimated the correlation between two individuals as described by[14]. We identified components that reflected genome-wide structure in two ways. First, we created two subsets of the data containing SNPs from the odd- and even-numbered chromosomes, repeated the PCA on each of these, and inspected scatter plots of pairs of components between the two subsets of the data. A component which is due to a region of linkage disequilibrium on a chromosome (as opposed to genome-wide structure) will appear only when analysing the data set containing SNPs from that chromosome. Second, we computed the score of every SNP on the components. For a component that is due to a region of linkage disequilibrium, there will be a spike of high SNP scores only in that region. To minimize the contribution from regions of extensive strong linkage disequilibrium, the correlation estimates were based on a subset of 197,175 SNPs that were spaced at least 0.001 cM apart (HapMap estimates) and specifically excluded the MHC region.

To assess the level of over-dispersion in each collection we first created a very clean set of data to ameliorate the effects of over-dispersion due to calling problems and missing data. In addition to the main filters described above, we filtered out all SNPs that had a clear genotype-calling problem revealed by visual inspection, SNPs with a study-wide missing data rate >1% and SNPs with study-wide minor allele frequency <1%. Around 360,000 SNPs passed these filters. Estimates of λ were calculated using an estimator based on the median test statistic[15]. Estimates of λ were also calculated from tests that conditioned on the scores for each individual along the two estimated principal components described above. The tests (1 d.f. and 2 d.f.) were carried out by including the scores as additional covariates in a logistic regression model fit.

Bayes factors. The box in the main text makes the point that understanding the strength of evidence conveyed by a particular P value also requires knowledge of power. In contrast, the Bayes factor (BF) provides a single measure of the strength of the evidence for an association, and we report these in addition to P values (Supplementary Table 14). As for power, calculation of Bayes factors requires assumptions about effect sizes. The assumptions underlying our calculations are given below and in Supplementary Information.

There is broad agreement between the way in which P values and our Bayes factors rank SNPs, except for SNPs with low MAFs (Supplementary Fig. 22). This is intuitive: unless one believed, a priori, that rare causative SNPs have substantially larger effect sizes, there will be reduced power for these SNPs and hence weaker evidence for association than for common SNPs with the same P value.

One perspective on GWAs is that in practice they will be used to prioritize SNPs for further study or additional typing. In addition to BFs providing a single quantity that can be directly compared between SNPs, it is also straightforward for investigators to give different a priori weights to different classes of SNPs, such as non-synonymous (ns)SNPs, genic SNPs, SNPs in highly conserved regions, or SNPs in linkage disequilibrium with many (or few) other SNPs.

We now describe calculation of the Bayes factors. We use M_0 to denote a model of no association, M_1 for a model with an additive effect on the log-odds scale and M_2 for a general 3 parameter model of association. At each SNP we calculate two Bayes factors: one for the additive model versus the null model, BF_1, and one for the general model versus the null model, BF_2. That is,

$$BF_1 = \frac{\Pr(Data|M_1)}{\Pr(Data|M_0)},\ BF_2 = \frac{\Pr(Data|M_2)}{\Pr(Data|M_0)},$$

where $\Pr(Data|M_i) = \int \Pr(Data|\theta_i, M_i)\Pr(\theta_i|M_i)d\theta$, where θ denotes the parameters for the model. For all 3 models we use a logistic regression model for the likelihood $\Pr(Data|\theta_i, M_i)$ where the log-odds for individual i is equal to μ for model M_0, $\mu+\gamma Z_i$ for model M_1 and $\mu+\gamma I(Z_i=1)+\phi(2\gamma I(Z_i=2))$ for model M_2. Z_i is the genotype (coded 0, 1 and 2) for individual i and $I(Z_i=m)$ is the indicator function that individual i has the genotype coded as m. For each model we choose the priors on the parameters, $\Pr(\theta_i|M_i)$, to reflect our belief about the likely effect sizes underlying complex trait loci.

The parameter γ in models M_1 and M_2 is the increase in log-odds of disease for every copy of the allele coded as 1, and e^{γ} is the additive model odds ratio. For both models we use a $N(0, 0.2)$ prior on γ. This prior puts probability 0.31 on

odds ratios above 1.2 or below 0.8, and probability 0.02 on odds ratios above 1.5 or below 0.5. The parameter μ in all three models represents the baseline odds of disease. In a case-control design the numbers of cases in the sample have been elevated artificially, which will have a large effect on likely values of μ. Our prior beliefs about the baseline risk of disease must take this into account. For all three models we have used a $N(0, 1)$ for μ and have found that the resulting Bayes factors are relatively insensitive to choice of priors for this parameter as long as the same prior is used for the two models being compared. The parameter ϕ in model M_2 represents a recessive effect over and above an additive effect. We use a $N(1, 1)$ prior for ϕ. Combined with the prior on γ, this results in a prior probability of 0.25 on the odds ratios above 1.5 and below 0.5 for the genotype coded as 2. In addition, we note that the evaluation of the Bayes factors will depend on the way the alleles at the SNP have been coded 0 and 1. To account for this we average over the two possible codings of each SNP with equal weight. A fuller description of the priors used can be found in Supplementary Information.

Sex-differentiated tests. We examined the possibility of differential genetic effects in males and females by reapplying the two single-locus analyses (trend test and genotypic test) separately in males and females and combining the results (simply adding the chi-squared statistics for the male and female analyses, and comparing with the 2 d.f. or 4 d.f. null hypothesis; results are shown in Supplementary Table 15). We refer to this as a sex-differentiated test. This test is sensitive to association that is of a different magnitude and/or direction in the two sexes, although it is less powerful than the simple test when the effect size does not vary with sex.

X Chromosome analysis. For several reasons the X chromosome needs to be treated differently from the autosomes (note that the Affymetrix chip used does not assay the Y chromosome). First, samples sizes and hence power are different from the autosomes (only one copy of X in males). Also, because the effective population size on the X chromosome is smaller than the autosomes, linkage disequilibrium extends further. And unlike the autosomes, there are choices in how to implement even single locus analyses: these relate to the relative weight to be given to males and females in comparisons between cases and controls.

For autosomal SNPs, the 1 d.f. trend test statistic is calculated by dividing the square of the difference between means of the SNP genotypes (scored 0, 1, 2) between cases and controls by an estimate of its variance. The variance estimate used is an empirical estimate that does not assume Hardy–Weinberg equilibrium. The numerator can also be represented as the squared difference in allele frequencies between cases and controls, as in the allele counting test. At first sight, a natural generalization of this test to deal with SNPs on the X chromosome would involve comparing allele frequencies, by allele counting, but using a variance estimate which does not assume Hardy–Weinberg equilibrium in females. However, we took the view that, because most loci on the X chromosome are subject to X chromosome inactivation, it is more logical to treat males as if they were homozygous females. Thus we score female genotypes 0, 1 or 2 and male genotypes 0 or 2, comparing mean scores of cases and controls as before. The variance estimate allows for the different variance of male and female contributions and does not assume Hardy–Weinberg equilibrium in females.

A stratified version of the test is constructed using the same principles by which the trend test is extended to the Mantel extension test; a score that contrasts cases and controls is computed for each stratum together with its variance; these are then summed over strata. The final test is the squared total score divided by the total variance. To extend these tests to a 2 d.f. test, we add a score that compares heterozygosity between cases and controls. Clearly, only females contribute to this component. Results of these analyses of X chromosome SNPs are shown in Supplementary Table 16.

Multilocus analysis. We use (1) the genotype data of this study, (2) the HapMap data, and (3) a population genetics model, to simulate genotypes at the HapMap SNPs that are not on the Affymetrix 500K chip. Informally, we determine which haplotypes are present in each individual in a region, and then use HapMap to 'fill in' these haplotypes at untyped SNPs (see below for details). These 'in silico' genotypes are then tested for association with the disease as before. This powerful multilocus tool for association studies[143] has the advantage of using information from all markers in linkage disequilibrium with an untyped SNP, but in a way that decreases with genetic distance. Our imputation method was applied to individuals passing project filters, and used markers which passed the project filters and in addition had MAF $>$ 1%. As a validation we compared our imputed genotypes for 58C individuals with genotypes obtained on an Illumina platform for 10,180 SNPs that are polymorphic in CEU HapMap samples. At these SNPs, for imputed genotypes with posterior call probabilities above 0.95, there was 98.4% agreement with the Illumina genotypes.

In our association analyses we imputed genotypes at 2,139,483 HapMap SNPs, and tested these for association with each disease using the trend test or the genotypic test. We included the results from imputed SNPs in the signal plots (Fig. 5) because they are useful in (1) assessing signal strength within a region; (2) providing a wider range of SNPs for follow up; and (3) indicating possible locations for the causal variant. For example in the case of *TCF7L2* in T2D, there is a substantially stronger signal from rs7903146 than for any of the typed SNPs (see also Supplementary Fig. 12).

To be conservative, stringent quality control filters were applied to genomic regions where imputed SNPs (but not genotyped SNPs) were responsible for a strong signal for association. These were as follows: (1) any such region was required to contain more than one imputed SNP showing the required level of association with a MAF $>$ 2% and posterior probability for imputed genotypes averaged across the SNP $>$0.95 (empirical studies showed imputation at low MAF SNPs more prone to error); (2) all cluster plots for genotyped SNPs within 0.3 cM (from HapMap Phase II estimated recombination rates) were checked and where there was evidence of any mis-calling the region was rejected (the major problem with imputation arises around SNPs with genotype calling errors); and (3) if there was no genotyped SNP with a P value $< 10^{-4}$ for association on either trend or genotypic test, the region was rejected. Note that accuracy of imputation with these filters applied will be larger than the figure of 98.4% reported above.

We use $H=\{H_1,\ldots, H_N\}$ to denote a set of N known haplotypes where $H_i=\{H_{i1},\ldots, H_{iL}\}$ is an individual haplotype and L is the number of SNP loci. In practice, we set H to be the 120 CEU haplotypes estimated as part of the HapMap project owing to the expected similarity in haplotype structure between the CEU and UK populations. We let $G=\{G_1,\ldots, G_k\}$ denote the genotype data on the K individuals in the study where $G_i=\{G_{i1},\ldots, G_{iL}\}$ and $G_{ij}\in\{0, 1, 2, \text{missing}\}$. In this setting, the majority of SNPs will have entirely missing genotypes, because the Affymetrix 500K chip has approximately 1/6th of the number of SNPs in the Phase II HapMap. The missing genotypes are imputed by modelling the distribution of each individual's genotype vector G_i conditional on the known set of haplotypes H, $\Pr(G_i|H)$. Our model for each individual's genotype vector is a Hidden Markov Model in which the hidden states are a sequence of pairs of the N known haplotypes in the set H. That is,

$$\Pr(G_i|H)=\sum_{Z_i^{(1)},\,Z_i^{(2)}}\Pr\left(G_i\middle|Z_i^{(1)},\,Z_i^{(2)},\,H\right)\Pr\left(Z_i^{(1)},\,Z_i^{(2)}\right),$$

where $Z_i^{(1)}=\left\{Z_{i1}^{(1)},\ldots,Z_{iL}^{(1)}\right\}$ and $Z_i^{(2)}=\left\{Z_{i1}^{(2)},\ldots,Z_{iL}^{(2)}\right\}$ are the two sequences of copying states at the L sites and $Z_{il}^{(j)}\in\{1,\ldots,N\}$. Here, $\Pr\left(Z_i^{(1)},\,Z_i^{(2)}\right)$ defines our prior probability on how the sequences of copying states change along the sequence and $\Pr\left(G_i\middle|Z_i^{(1)},\,Z_i^{(2)},\,H\right)$ models how the observed genotypes will be close to but not exactly the same as the haplotypes being copied. The precise form of these terms (described in ref. 142) are based on an approximate population genetics model that makes direct use of the recently estimated fine-scale recombination map across the genome[142,143]. At each of the missing genotypes in the study, we use this model to calculate probabilities for the three possible genotypes. At each imputed SNP, we used these probabilities to calculate the 2×3 table of expected genotype counts for cases and controls and used these counts to carry out a standard test of association.

Disease models. To test for deviations from additivity (in log-odds) at a locus we fit a logistic regression model using the function glm in the statistical software R (http://www.r-project.org/). For each region we considered the most significant SNP and compared an additive model to a general 2-d.f. model by fitting a model with an additive sub-model nested in a general model. The additive effect was modelled by a variable encoded 0, 1, or 2 for the effect at the three genotypes and a second term for a general model was included by a variable encoded 1 for heterozygotes and 0 otherwise. We rejected an additive model if the second term was significant and then compared a dominant or recessive model to a general model. For the pairwise interaction analysis, we fixed the marginal model at each locus on the basis of the single locus analysis. We compared the two locus model with these marginals and no interaction terms with a larger model including interactions. This larger interaction model has 1, 2, or 4 additional parameters depending on whether both marginal models are additive, one is additive and one general, or both general.

Software. Several software packages were developed within the WTCCC for data analysis, data management and simulation studies. We found it necessary to normalize the Affymetrix probe intensity data to minimize chip-to-chip variability. A C++ program was written to carry out this normalization efficiently. To obtain a copy of the software please email Hin-Tak Leung at hin-tak.leung@cimr.cam.ac.uk.

We developed a new genotype calling algorithm, CHIAMO, implemented in C++. CHIAMO uses a hierarchical statistical model, which allows it to simultaneously call genotypes at all data samples. To obtain a copy of the software please email J. L. Marchini at marchini@stats.ox.ac.uk.

To perform genome-wide association analysis we developed two software packages: snpMatrix and SNPTEST. snpMatrix is an R package and is freely available from http://www-gene.cimr.cam.ac.uk/clayton/software/. Both quantitative and qualitative phenotypes can by analysed using snpMatrix and flexible association testing functions are provided that control for potential confounding by quantitative and qualitative covariates. SNPTEST is a standalone C++ program that implements both frequentist tests and bayesian analysis of association and allows the user to include quantitative or qualitative covariates. This program works directly with the output of CHIAMO and IMPUTE (see below). To obtain a copy of the software please email J. L. Marchini at marchini@stats.ox.ac.uk.

Genotypes at SNPs that are in HapMap but not on the Affymetrix 500K chip were imputed using the C++ program IMPUTE, which makes use of genotype information at neighbouring SNPs. To obtain a copy of the software please email J. L. Marchini at marchini@stats.ox.ac.uk.

119. Spitzer, R. L., Endicott, J. & Robins, E. Research diagnostic criteria: rationale and reliability. *Arch. Gen. Psychiatry* **35,** 773–782 (1978).
120. Wing, J. K. B. T. *et al.* SCAN. Schedules for Clinical Assessment in Neuropsychiatry. *Arch. Gen. Psychiatry* **47,** 589–593 (1990).
121. Craddock, M. *et al.* Concurrent validity of the OPCRIT diagnostic system. Comparison of OPCRIT diagnoses with consensus best-estimate lifetime diagnoses. *Br. J. Psychiatry* **169,** 58–63 (1996).
122. McGuffin, P., Farmer, A. & Harvey, I. A polydiagnostic application of operational criteria in studies of psychotic illness. Development and reliability of the OPCRIT system. *Arch. Gen. Psychiatry* **48,** 764–770 (1991).
123. Green, E. K. *et al.* Operation of the schizophrenia susceptibility gene, neuregulin 1, across traditional diagnostic boundaries to increase risk for bipolar disorder. *Arch. Gen. Psychiatry* **62,** 642–648 (2005).
124. Green, E. K. *et al.* Genetic variation of brain-derived neurotrophic factor (BDNF) in bipolar disorder: case-control study of over 3000 individuals from the UK. *Br. J. Psychiatry* **188,** 21–25 (2006).
125. Samani, N. J. *et al.* A genomewide linkage study of 1,933 families affected by premature coronary artery disease: The British Heart Foundation (BHF) Family Heart Study. *Am. J. Hum. Genet.* **77,** 1011–1020 (2005).
126. Lennard-Jones, J. E. Classification of inflammatory bowel disease. *Scand. J. Gastroenterol.* (Suppl.) **170**, 2–6; discussion 6–9 (1989).
127. Arnett, F. C. *et al.* The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum.* **31,** 315–324 (1988).
128. MacGregor, A. J., Bamber, S. & Silman, A. J. A comparison of the performance of different methods of disease classification for rheumatoid arthritis. Results of an analysis from a nationwide twin study. *J. Rheumatol.* **21,** 1420–1426 (1994).
129. Worthington, J. *et al.* The Arthritis and Rheumatism Council's National Repository of Family Material: pedigrees from the first 100 rheumatoid arthritis families containing affected sibling pairs. *Br. J. Rheumatol.* **33,** 970–976 (1994).
130. Symmons, D. P., Barrett, E. M., Bankhead, C. R., Scott, D. G. & Silman, A. J. The incidence of rheumatoid arthritis in the United Kingdom: results from the Norfolk Arthritis Register. *Br. J. Rheumatol.* **33,** 735–739 (1994).
131. Smyth, D. *et al.* Replication of an association between the lymphoid tyrosine phosphatase locus (*LYP/PTPN22*) with type 1 diabetes, and evidence for its role as a general autoimmunity locus. *Diabetes* **53,** 3020–3023 (2004).
132. Wiltshire, S. *et al.* A genomewide scan for loci predisposing to type 2 diabetes in a U.K. population (the Diabetes UK Warren 2 Repository): analysis of 573 pedigrees provides independent replication of a susceptibility locus on chromosome 1q. *Am. J. Hum. Genet.* **69,** 553–569 (2001).
133. Frayling, T. M. *et al.* Parent–offspring trios: a resource to facilitate the identification of type 2 diabetes genes. *Diabetes* **48,** 2475–2479 (1999).
134. Groves, C. J. *et al.* Association analysis of 6,736 U.K. subjects provides replication and confirms *TCF7L2* as a type 2 diabetes susceptibility gene with a substantial effect on individual risk. *Diabetes* **55,** 2640–2644 (2006).
135. Power, C. & Elliott, J. Cohort profile: 1958 British birth cohort (National Child Development Study). *Int. J. Epidemiol.* **35,** 34–41 (2006).
136. Strachan, D. P. *et al.* Lifecourse influences on health among British adults: Effects of region of residence in childhood and adulthood. *Int. J. Epidemiol.* Advance online publication, doi:10.1093/ije/dyl309 (25 January 2007).
137. Matsuzaki, H. *et al.* Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. 1, 104–105. *Nat Methods* **1,** 104–105 (2004).
138. Di, X. *et al.* Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays. *Bioinformatics* **21,** 1958–1963 (2005).
139. Rabbee, N. & Speed, T. A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics* **22,** 7–12 (2006).
140. Affymetrix. in Technical Report (2006).
141. Stirling, W. D. Enhancements to Aid Interpretation of Probability Plots. *Statistician* **31,** 211–220 (1982).
142. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165,** 2213–2233 (2003).
143. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies via imputation of genotypes. *Nature Genet.* doi:10.1038/ng2088 (in the press).