Genetic Epidemiology 36 : 785–796 (2012)

Genome-Wide Association Analysis of Imputed Rare Variants: Application to Seven Common Complex Diseases

Reedik Mägi,[1] Jennifer L. Asimit,[2] Aaron G. Day-Williams,[2] Eleftheria Zeggini,[2] and Andrew P. Morris[3]*

[1] *Estonian Genome Centre, University of Tartu, Tartu, Estonia*
[2] *Wellcome Trust Sanger Institute, Hinxton, United Kingdom*
[3] *Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom*

Genome-wide association studies have been successful in identifying loci contributing effects to a range of complex human traits. The majority of reproducible associations within these loci are with common variants, each of modest effect, which together explain only a small proportion of heritability. It has been suggested that much of the unexplained genetic component of complex traits can thus be attributed to rare variation. However, genome-wide association study genotyping chips have been designed primarily to capture common variation, and thus are underpowered to detect the effects of rare variants. Nevertheless, we demonstrate here, by simulation, that imputation from an existing scaffold of genome-wide genotype data up to high-density reference panels has the potential to identify rare variant associations with complex traits, without the need for costly re-sequencing experiments. By application of this approach to genome-wide association studies of seven common complex diseases, imputed up to publicly available reference panels, we identify genome-wide significant evidence of rare variant association in *PRDM10* with coronary artery disease and multiple genes in the major histocompatibility complex (MHC) with type 1 diabetes. The results of our analyses highlight that genome-wide association studies have the potential to offer an exciting opportunity for gene discovery through association with rare variants, conceivably leading to substantial advancements in our understanding of the genetic architecture underlying complex human traits. *Genet. Epidemiol.* 36:785–796, 2012.

Key words: genome-wide association study; rare variants; imputation

Supporting Information is available in the online issue at wileyonlinelibrary.com.
Contract grant sponsor: Development Fund of the University of Tartu; Contract grant sponsor: Wellcome Trust; Contract grant numbers: WT098051; WT081682; WT076113; WT090532.
*Correspondence to: Andrew P. Morris, Genetic and Genomic Epidemiology Unit, Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. E-mail: amorris@well.ox.ac.uk
Received 16 April 2012; Revised 23 July 2012; Accepted 27 July 2012
Published online 5 September 2012 in Wiley Online Library (wileyonlinelibrary.com/journal/gepi).
DOI: 10.1002/gepi.21675

INTRODUCTION

There has been much recent debate as to the role of rare genetic variation, defined here to have a minor allele frequency (MAF) of less than 1%, in explaining the 'missing heritability' of complex traits [Dickson et al., 2010; Frazer et al., 2009; Yang et al., 2010]. Rare variants are likely to have originated from founder effects in the last 20 generations, and thus are more likely to be population specific [Bodmer and Bonilla, 2008]. They are also likely to have larger effects on complex traits than common variants, consistent with the expectation that they will have been subject to purifying selection after recent expansion of the human population [Pritchard, 2001]. However, these effects are unlikely to be sufficiently large to be detected through association testing with individual rare variants. Statistical methods have thus focussed on the aggregation of the effects of all rare variants within the same exon, gene or pathway, potentially weighting according to annotation or MAF [Han and Pan, 2010; Hoffman et al., 2010; Li et al., 2010; Li and Leal, 2008; Madsen and Browning, 2009; Morgenthaler and Thilly, 2007; Morris and Zeggini, 2010; Neale et al., 2011; Price et al., 2010; Wu et al., 2011; Zawistowski et al., 2010]. Using these methods, multiple rare variants have been demonstrated to be associated with a variety of complex traits including low- and high-density lipoprotein [Cohen et al., 2006; Romeo et al., 2007], body mass index [Ahituv et al., 2007] and blood pressure [Ji et al., 2008].

The most comprehensive approach to characterising the contribution of rare variants to the genetic component of complex traits is through large-scale, next-generation re-sequencing studies [Metzker, 2010]. Despite improvements in the throughput and efficiency of these technologies, rare variant re-sequencing efforts on the scale of the whole genome still represent an infeasible financial undertaking for most research groups. Consequently, most rare variant studies have focussed on candidate genes, or more recently, the exome [Ng et al., 2010]. However, high-density reference panels obtained from whole-genome re-sequencing data are being released through the 1000 Genomes Project, providing a comprehensive catalogue of variation with MAF as low as 0.5%, as well as many

rarer variants, across a wide range of populations from different ethnic groups [The 1000 Genomes Project Consortium, 2010]. Such reference panels could be utilised to select rare variants for large-scale genotyping with custom designed arrays, potentially with priority given to variants with likely functional consequences in an effort to reduce costs, such as the Illumina Infinium HumanExome BeadChip. Conversely, genome-wide association study (GWAS) genotyping products have been designed primarily to capture common genetic variation, and thus offer poor coverage of rare variants [Barrett and Cardon, 2006]. However, if samples have already been assayed by means of such a GWAS chip, imputation techniques [Marchini and Howie, 2010] can make use of this existing scaffold to predict genotypes at variants present on the higher density reference panel, incurring no additional cost, other than computation, although this is far from trivial.

Here, we formulate methodology for the detection of complex trait association with accumulations of minor alleles within genes, or some other functional unit, using data from directly typed and/or imputed rare variants. We report the results of simulations to investigate the power of alternative design strategies for assaying and characterising rare genetic variation to detect association with a quantitative trait in a 50 kb gene. Our study considers a simple model of rare variant association with the trait, and assesses the impact on power of the number of individuals present in the reference panel. We also present results of an application of our methodology to rare variant association analysis of seven common complex diseases undertaken by the Wellcome Trust Case Control Consortium (WTCCC) [The Wellcome Trust Case Control Consortium, 2007], using GWAS data imputed up to the Phase I 1000 Genomes Project reference panel (June 2011 interim release) [The 1000 Genomes Project Consortium, 2010].

METHODS

MODEL FORMULATION

We test for association of a complex trait with accumulations of minor alleles at N rare variants, here defined to have MAF less than 1%, within the same exon, gene or some other functional unit, in a sample of unrelated individuals. Let n_i denote the number of rare variants at which the ith individual has been successfully genotyped in the functional unit. Furthermore, let G_{ij} denote the genotype of this individual at the jth rare variant, coded as 0 for the common homozygote, and 1 otherwise. We can model the phenotype, y_i, of this individual in a generalised linear regression framework as a function of the proportion of rare variants at which they carry at least one minor allele [Morris and Zeggini, 2010], given by

$$\rho_i = \sum_{j=1}^{N} G_{ij}/n_i.$$

Specifically, $E(y_i) = g^{-1}(\alpha + \beta\rho_i)$, where g is the link function and β is the expected increase in the phenotype for an individual carrying a full complement of minor alleles at rare variants in the functional unit compared to an individual carry none. It follows that β/N is the increase in the expected phenotype of an individual for each rare variant at which they carry a minor allele. The likelihood contribution of the ith individual, $f(y_i|\alpha,\ \beta,\ \rho_i)$, is weighted by n_i to allow for differential call rates between samples. We can thus construct a likelihood ratio test by comparing the maximised weighted likelihoods of two models via analysis of deviance: (i) the null model where $\beta = 0$; and (ii) the alternative model for which β is unconstrained. The resulting test statistic has an approximate χ^2 distribution with one degree of freedom. The flexibility of the generalised linear regression framework allows for generalisation of this approach to take account of non-genetic risk factors as covariates.

It is straightforward to accommodate imputed rare variants in the functional unit within the generalised linear regression framework by considering the posterior distribution of genotype calls. Specifically, we replace missing genotypes (at typed variants, or at untyped variants in the imputation reference panel) by their expectation, $E(G_{ij}) = 1 - h_{ij}$, where h_{ij} is the posterior probability of a common homozygote call in the ith individual at the jth rare variant. The posterior probabilities, h_{ij}, of imputed genotype calls are easily recovered from standard imputation software such as IMPUTEv2 [Howie et al., 2009] and BEAGLE [Browning and Browning, 2009].

The methodology described above has been implemented in the GRANVIL software, and is freely available for download (http://www.well.ox.ac.uk/GRANVIL). The open-source software has been designed to efficiently handle analysis of directly genotyped and imputed rare variants on a genome-wide scale, and can accommodate both quantitative traits and binary phenotypes in a generalised linear modelling framework, as described above. The user supplies a file containing the boundaries of each functional unit to be analysed, together with SNPTEST format sample and genotype files, and specifies the rare variant MAF threshold. GRANVIL is distributed with scripts to generate graphical summaries of GWAS rare variant analysis results, including quantile-quantile and Manhattan plots.

SIMULATION STUDY

We have performed simulations to evaluate the relative performance of different design strategies to identify quantitative trait association with rare variants in a 50 kb gene. We have considered an analysis cohort of 2,000 individuals, and a reference panel ascertained from the same population. We have compared the power of GRANVIL in the following scenarios: (i) direct re-sequencing of the analysis cohort; (ii) direct genotyping of the analysis cohort for all rare variants present in the reference panel; (iii) direct genotyping of the analysis cohort for variants on a GWAS chip with the same characteristics as the Illumina Human660W-Quad BeadArray in terms of density and MAF; and (iv) direct genotyping of the analysis cohort for variants on the GWAS chip, supplemented with imputation of untyped variants present in the reference panel using IMPUTEv2 [Howie et al., 2009].

We have considered a simple underlying model for the association of the trait with multiple rare causal variants within the same gene. We assumed that the expected trait value of each individual is increased by the presence of a minor allele at any causal variant. The trait association model was then parameterised in terms of: (i) the maximum MAF, δ, of any individual causal variant; (ii) the total MAF, Q, of all causal variants in the region; and (iii) their joint contribution to the overall trait variance, expressed as $100\lambda\%$. Here, we considered reference panels of $R = 120$, $R = 500$ and $R = 4{,}000$ individuals. The number of

individuals was chosen to represent a range of reference panels incorporating those available from the 1000 Genomes Project (pilot release), through to those we might expect from future large-scale deep re-sequencing efforts, such as the UK10K initiative (http://www.uk10k.org/). For each model, we generated 500 replicates of data as follows.

(1) Generate an ancestral recombination graph [Griffiths and Marjoram, 2007] for a population of 40,000 haplotypes from a realisation of the coalescent process with recombination, obtained using the MS software [Hudson, 2002]. We assumed a mutation rate of 10^{-8} per base (in each generation) and a uniform recombination rate of 1 cM per Mb, for an effective population size of 10,000 individuals. In total, we simulated a region of 1,050 kb, including a 50 kb gene and 500 kb up- and down-stream to allow for an imputation buffer to improve accuracy by avoiding edge effects and taking advantage of the expected long-range linkage disequilibrium (LD) with rare variants [The International HapMap Consortium, 2007].
(2) Calculate the MAF at each variant across the 50 kb gene in the population of 40,000 chromosomes, denoted by q_j for the jth variant. Select a variant as causal from amongst those with MAF $q_j < \delta$, at random. Continue selecting causal variants in this way, without replacement, until the total MAF of all causal variants is Q.
(3) Select a random sample of 4,000 chromosomes from the population, paired together to form the analysis cohort. Determine the number of minor alleles carried by the ith individual across all causal variants in the 50 kb gene, denoted by m_i. The phenotype, y_i, of the ith individual is then simulated from a Gaussian $N(\mu_i,\sigma)$ distribution, where σ is determined by the spectrum of causal variants and their joint contribution, λ, to the overall trait variance, and $\mu_i = 1$ if $m_i > 0$, and 0 otherwise. Full details of the derivation of the residual trait variance, σ^2, are provided in the Appendix.
(4) Select a random sample of $2R$ chromosomes from the remainder of the population to be haplotypes in the reference panel. Assuming no genotyping or phasing errors in the reference panel, record the haplotype of each of these chromosomes across all variants in the 1,050 kb region.
(5) Begin by considering the strategy in which the analysis cohort has been directly re-sequenced in the 50 kb gene. Assuming no sequencing errors, record the genotype of each individual at each variant with MAF < 1% in the analysis cohort. Test for association of the quantitative trait with an accumulation of minor alleles at these variants using GRANVIL, and record the P-value, denoted by p_{SEQ}.
(6) Then consider the scenario in which the analysis cohort has been directly genotyped for all variants in the 50 kb gene which are present in the reference panel. Assuming no genotyping errors, record the genotype of each individual at each variant present in the reference panel with MAF < 1% in the analysis cohort. Test for association of the quantitative trait with an accumulation of minor alleles at these variants using GRANVIL, and record the P-value, denoted by p_{GEN}.
(7) Next consider the scenario in which the analysis cohort has been directly genotyped only for variants on a GWAS chip. Select a random 1,050 kb region of the genome, and determine the number of variants, n_{GWAS}, present on the Illumina Human660W-Quad BeadArray in that interval. Select n_{GWAS} variants at random and without replacement, with ascertainment probability $\phi_j = 4q_j(1-q_j)/T$, as present on the chip, where $T = \sum_j q_j(1-q_j)$. This probability density incorporates the strong bias towards common variants on GWAS chips, generating an approximately uniform distribution of MAF [Anderson et al., 2008]. Assuming no genotyping errors, record the genotype of each individual at each variant within the 50 kb gene with MAF < 1% in the analysis cohort. Test for association of the quantitative trait with an accumulation of minor alleles at these variants using GRANVIL, and record the P-value, denoted by p_{GWAS}.
(8) Finally, consider the scenario in which the analysis cohort has been directly genotyped only for variants on the GWAS chip, but are subsequently supplemented with imputation of untyped variants present in the reference panel. Assuming no genotyping errors, record the genotype of each individual at each variant across the 1,050 kb region, irrespective of MAF. Impute the genotype of each individual in the analysis cohort at each variant present in the reference panel in the 50 kb gene using IMPUTEv2 [Howie et al., 2009], assuming an effective population size of 10,000 individuals and a buffer region of 500 kb. Test for association of the quantitative trait with an accumulation of minor alleles at directly genotyped and imputed variants within the 50 kb gene with MAF < 1% in the analysis cohort and 'info score' greater than 0.4 using GRANVIL, and record the P-value, denoted by p_{IMP}.

Over all simulated data sets, we calculated the power of each design strategy at a nominal 5% significance level as the proportion of replicates for which the corresponding P-value is less than 0.05. We also calculated the mean numbers of rare variants in the 50 kb gene: (i) in the population of 40,000 chromosomes; (ii) identified through direct re-sequencing of the analysis cohort; (iii) present on the reference panel and identified through direct genotyping of the analysis cohort; (iv) present on the GWAS chip; (v) present on the reference panel and well imputed (info score greater than 0.4) in the analysis cohort.

One potential limitation of our simulation study is the assumption of no sequencing errors, perfect phasing of the reference panel, and no missing or miscalled genotypes. These errors might be expected to be most detrimental to our proposed imputation strategy since this process requires an accurate GWAS scaffold and phased reference panel. To address this issue, we have also performed simulations to assess the robustness of our results to re-sequencing errors and missing or miscalled genotypes. We introduced a simple model of errors [Hao et al., 2004] by randomly swapping the base call (ancestral or mutant) of each individual at each variant with probability ε. Here, we considered $\varepsilon = 0\%$ (no errors), $\varepsilon = 0.1\%$ (error rate of $\sim$0.2% in genotype calls) and $\varepsilon = 0.2\%$ (error rate of $\sim$0.4% in genotype calls). We evaluated the impact of call rate by removing genotypes, at random, with probability κ, where here we considered $\kappa = 0\%$ (no missing genotype data) and $\kappa = 1\%$. Errors were introduced first into the reference panel (Step 4), and then into the analysis cohort (Steps 5, 6, 7 and 8). Missing genotype data were then introduced at random (Steps 6, 7 and 8).

APPLICATION TO IMPUTED RARE VARIANT GWAS OF SEVEN COMMON COMPLEX DISEASES

We considered 14,000 cases of seven common complex diseases (bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes and type 2 diabetes) and 3,000 shared controls from the WTCCC [The Wellcome Trust Case Control Consortium, 2007]. Samples were ascertained from the United Kingdom and genotyped using the Affymetrix GeneChip 500K Mapping Array Set, which incorporates 500,568 single nucleotide polymorphisms (SNPs) genome-wide. We utilised the same quality control (QC) filters employed by the WTCCC to exclude samples and SNPs from the analysis, full details of which are presented in the description of the experiment [The Wellcome Trust Case Control Consortium, 2007]. In brief, samples were excluded on the basis of low call rate, outlying genome-wide heterozygosity, discrepancies in WTCCC and external identifying information, non-European ancestry, duplication and apparent relatedness. SNPs were excluded on the basis of low call rate, extreme deviation from Hardy-Weinberg equilibrium (HWE), differential allele or genotype frequencies between the two control cohorts and manual visual inspection of genotype calls in cluster plots.

To allow for fine-scale population structure, which may have greater impact on rare variant association signals than common SNPs because of recent founder effects [Bodmer and Bonilla, 2008], we constructed principal components to represent axes of genetic variation within the UK. We applied EIGENSTRAT [Price et al., 2006] to a subset of high-quality LD-pruned SNPs ($r^2 < 0.2$) with MAF of at least 5%, and projected samples onto principal components demonstrating clear separation between 12 UK regions of residence [The Wellcome Trust Case Control Consortium, 2007].

We imputed the high-quality samples up to the Phase I 1000 Genomes Project reference panel (June 2011 interim release) consisting of 1,094 phased individuals from multiple ancestry groups [The 1000 Genomes Project Consortium, 2010]. We removed SNPs with MAF < 1% from the GWAS scaffold, prior to imputation, since we expect these variants are likely to be subject to higher genotyping errors, which may impact the downstream analysis. We performed imputation using IMPUTEv2 [Howie et al., 2009] with default parameter settings and sample pre-phasing, allowing a buffer region of 500 kb. Subsequently, we tested for association of each disease with 'high-quality' rare variants (MAF less than 1%, and IMPUTEv2 info score greater than 0.4) within genes using GRANVIL, adjusting for principal components as covariates to account for fine-scale UK population structure. Gene boundaries were defined from the UCSC human genome database (build 37).

RESULTS

SIMULATION STUDY

Figure 1 presents the power, at a nominal significance level of $P < 0.05$, to detect association with a quantitative trait, for each of the design strategies for assaying rare genetic variation in the gene. For these results, we assumed that multiple rare causal variants in the gene jointly contribute to 5% of the overall trait variation. The panels correspond to two specific trait association models: (A) the maximum MAF of any individual causal variant is 1%, and the total MAF of all causal variants is 5%; and (B) the maximum MAF of any individual causal variant is 0.5%, and the total MAF of all causal variants is 2%. Under the second of these models, we expect fewer rare variants within the gene to be causal, since the total MAF is lower. A higher proportion of non-causal variants within the gene would be expected to reduce power overall, irrespective of the design strategy and/or the number of individuals in the reference panel [Morris and Zeggini, 2010].

Our results highlight a number of general conclusions across these trait association models. As expected, the most powerful strategy to detect rare variant association is to resequence the analysis cohort. In the absence of sequencing errors, this 'gold-standard' strategy provides complete coverage of rare genetic variation in the gene within the analysis cohort. Nevertheless, genotyping the analysis cohort for all rare variants present in the reference panel generally results in a relatively small reduction of power, particularly for $R = 4{,}000$. We expect most of the rare variation in the analysis cohort to be captured by such large reference panels (Supporting Information Figure S1). Rare variants not captured by the reference panel (e.g., private mutations) are less likely to have a major impact on the joint contribution of causal variation in the gene under our simulation model, and thus would not be expected to lead to a dramatic reduction in power.

As previously reported [Morris and Zeggini, 2010], genotyping of the analysis cohort with the GWAS chip alone has minimal power to detect association because very few rare variants within the gene are assayed directly (Supporting Information Figure S1). However, imputation into this GWAS scaffold in the analysis cohort up to the density of the reference panel can lead to substantial gains in power to detect rare variant association within the gene. The extent of the increase in power depends crucially on the number of individuals in the reference panel, although the gains from $R = 500$ to $R = 4{,}000$ are not as great as from $R = 120$ to $R = 500$, particularly for an association model incorporating causal variants with MAF up to 1% (Figure 1). Reference panels with more individuals provide more comprehensive coverage of rare variation in the region (Supporting Information Figure S1), higher quality imputation, and thus greater improvements in power. Note that the relative power of imputation appears lower under trait association model (A), where the maximum MAF of any causal variant is lower than under model (B). This is not unexpected since the distribution of causal allele frequencies will be more skewed to the rarest variants under this model, which we anticipate to be most difficult to impute, irrespective of the size of the reference panel.

We also considered the impact of sequencing errors and missing and/or miscalled genotypes on the power of the four alternative strategies. As expected, the power of all strategies is decreased as the error rates and the frequency of missing genotypes increase (Supporting Information Figure S2). However, we are still able to recover much of the power of the gold-standard resequencing strategy through imputation of the analysis cohort from the GWAS scaffold, and we maintain considerable advantages over genotyping of the GWAS chip alone.

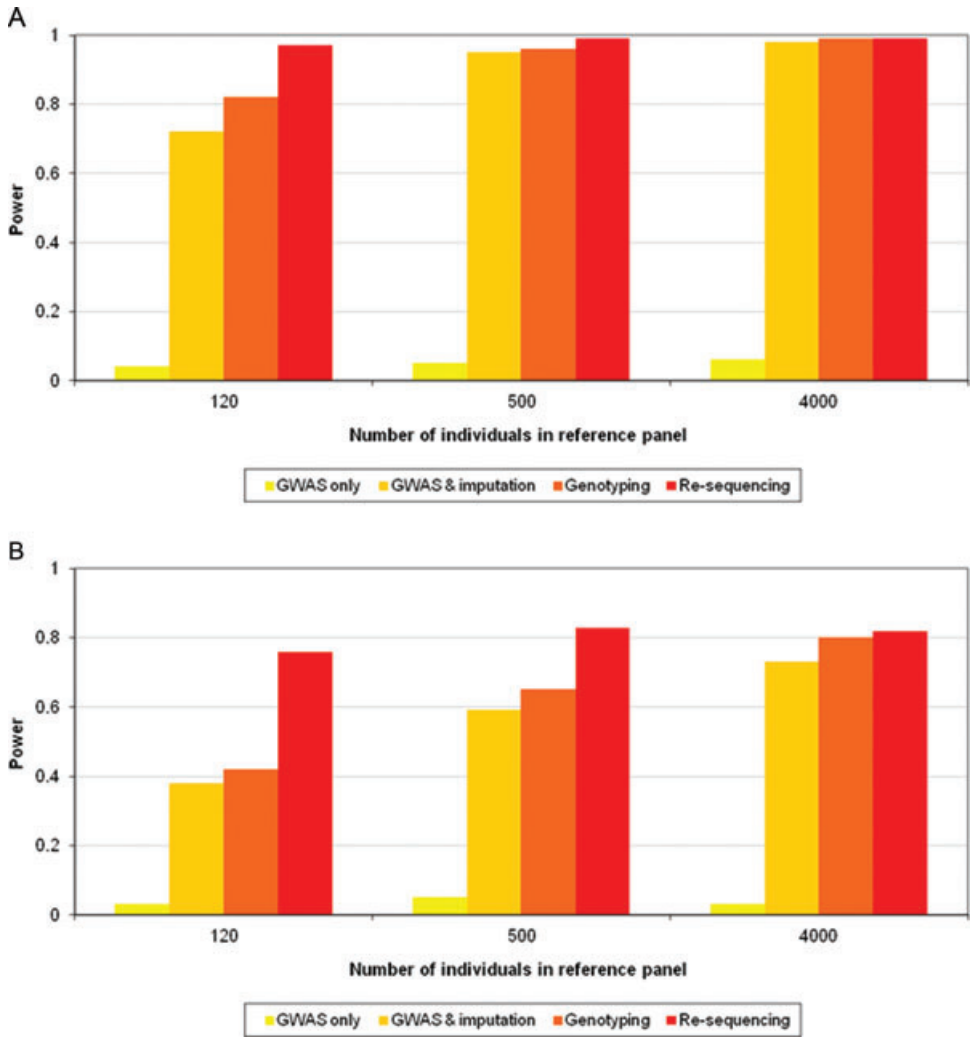

Fig. 1. Power, at a nominal significance level of $P < 0.05$, to detect association of an accumulation of minor alleles with a quantitative trait, for different strategies for assaying rare genetic variation in a 50 kb gene, as a function of the size of the reference panel. Multiple causal variants in the gene contribute jointly to 5% of the overall trait variation. The panels correspond to two specific trait association models: (A) the maximum MAF of any individual causal variant is 1%, and the total MAF of all causal variants is 5%; and (B) the maximum MAF of any individual causal variant is 0.5%, and the total MAF of all causal variants is 2%.

APPLICATION TO IMPUTED RARE VARIANT GWAS OF SEVEN COMMON COMPLEX DISEASES

A total of 13,241 cases and 2,938 controls from the WTCCC experiment passed sample QC filters (Supporting Information Table S1). Of the autosomal variants on the array, 456,868 passed SNP QC filters. We then applied EIGENSTRAT to an LD-pruned ($r^2 > 0.2$) set of 27,770 high-quality autosomal SNPs with MAF > 5% to construct 10 axes of genetic variation of UK population structure. By projecting samples onto the corresponding principal components, we observed that the first three axes of genetic variation were strongly associated with the region of residence of samples (Figure 2). The first principal component separated London and Scotland from the remainder of the United Kingdom, whilst the second and third principal components separated regions within the United Kingdom on a North-West to South-East axis. These three principal components were thus selected for adjustment of downstream association analyses to allow for fine-scale UK population structure.

After removal of variants with MAF < 1%, a total of 391,060 high-quality SNPs remained in the GWAS scaffold. A total of 8,239,134 rare variants were successfully imputed up to the Phase I 1000 Genomes Project reference panel (June 2011 interim release) [The 1000 Genomes Project Consortium, 2010] and were polymorphic in the WTCCC experiment. Of these, 5,383,228 (65.3%) had IMPUTEv2 info score of at least 0.4. Amongst these 'well-imputed' rare variants, the mean info score was 0.618, and 17.3% had info score greater than 0.8.

Figure 3 presents Manhattan plots to summarise the association of each disease with accumulations of minor alleles at well-imputed rare variants within genes, after correction for the three axes of genetic variation as covariates in the

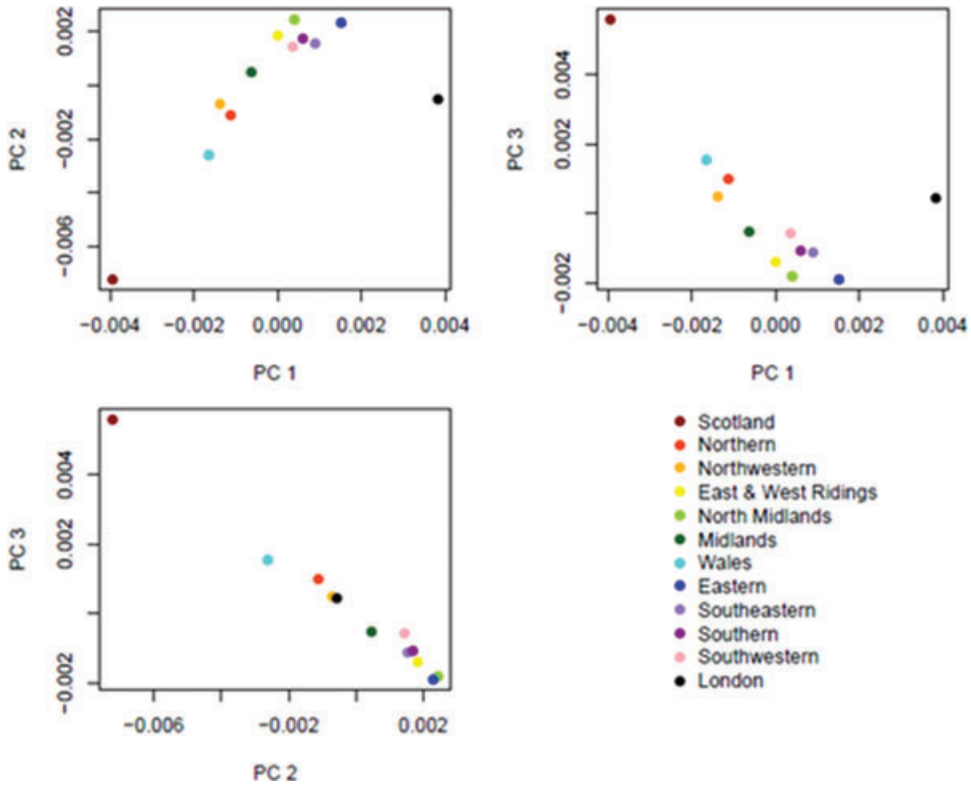

Fig. 2. Principal components representing axes of genetic variation demonstrating clear separation between 12 UK regions of residence. Each point represents the mean projection of samples from each UK region onto the first three axes of genetic variation.

logistic regression model. In these Manhattan plots, each point represents a gene (as defined by the UCSC human genome database), and those achieving genome-wide significance (Bonferroni correction for 30,000 genes, $P < 1.7 \times 10^{-6}$) are highlighted in red. There was no evidence of residual population structure, not accounted for by the three axes of genetic variation, with genomic control inflation factors [Devlin and Roeder, 1999] less than one for all seven diseases (Supporting Information Figure S3).

We observed genome-wide significant evidence of association of coronary artery disease with rare variants in the gene *PRDM10* ($P = 4.9 \times 10^{-8}$). The gene contained 122 well-imputed rare variants with mean MAF of 0.23%. Accumulations of minor alleles across these variants were associated with decreased risk of disease: odds ratio 0.828 (0.774–0.886) per minor allele. We also observed 10 genes with genome-wide significant evidence of rare variant association with type 1 diabetes, all located within the major histocompatibility complex (MHC) (Table I and Figure 4). The strongest signal of association was observed for *HLA-DRA* ($P = 2.0 \times 10^{-13}$), which has been previously implicated in susceptibility to type 1 diabetes [Nejentsev et al., 2007]. Accumulations of minor alleles at rare variants in nine of the MHC genes were associated with reduced risk of type 1 diabetes (Table I). The only gene demonstrating evidence of association of accumulations of minor alleles with increased risk of type 1 diabetes was *TNXA*, with odds ratio 2.346 (1.772–3.107) per minor allele.

Common SNPs in the MHC have been previously associated with the disease [Barrett et al., 2009; The Wellcome Trust Case Control Consortium, 2007], although fine-mapping of the underlying causal variant(s) has been hindered by the extensive LD across the region. We thus repeated our analyses in this region, testing for association of type 1 diabetes with rare variants within MHC genes after adjustment for the lead GWAS SNP (rs9268645) [Barrett et al., 2009], with genotypes coded by the number of minor alleles, included as an additional covariate in the logistic regression model (Table I and Figure 4). The common SNP could not fully explain rare variant associations of type 1 diabetes with any of the MHC genes, but dramatically reduced significance with *TNXA* ($P = 2.6 \times 10^{-9}$ before adjustment; $P = 2.4 \times 10^{-4}$ after adjustment). After adjustment, three additional MHC genes achieved genome-wide significant evidence of rare variant association with type 1 diabetes: *HLA-DMA* ($P = 1.1 \times 10^{-7}$), *SKIV2L* ($P = 2.6 \times 10^{-7}$) and *TNXB* ($P = 4.1 \times 10^{-7}$).

DISCUSSION

GWAS has been extremely successful in identifying genetic loci contributing effects to a wide range of complex human traits [Hindorff et al., 2009] including diseases such as type 2 diabetes [Voight et al., 2010] and Crohn's disease [Franke et al., 2010], and quantitative phenotypes such as body mass index [Speliotes et al., 2010] and height [Lango Allen et al., 2010]. However, despite the success of this approach, much of the genetic component of these traits remains, as yet, unexplained [Manolio et al., 2009]. Most of the confirmed associations within these loci are with common variants, of modest effect, which well-designed GWAS

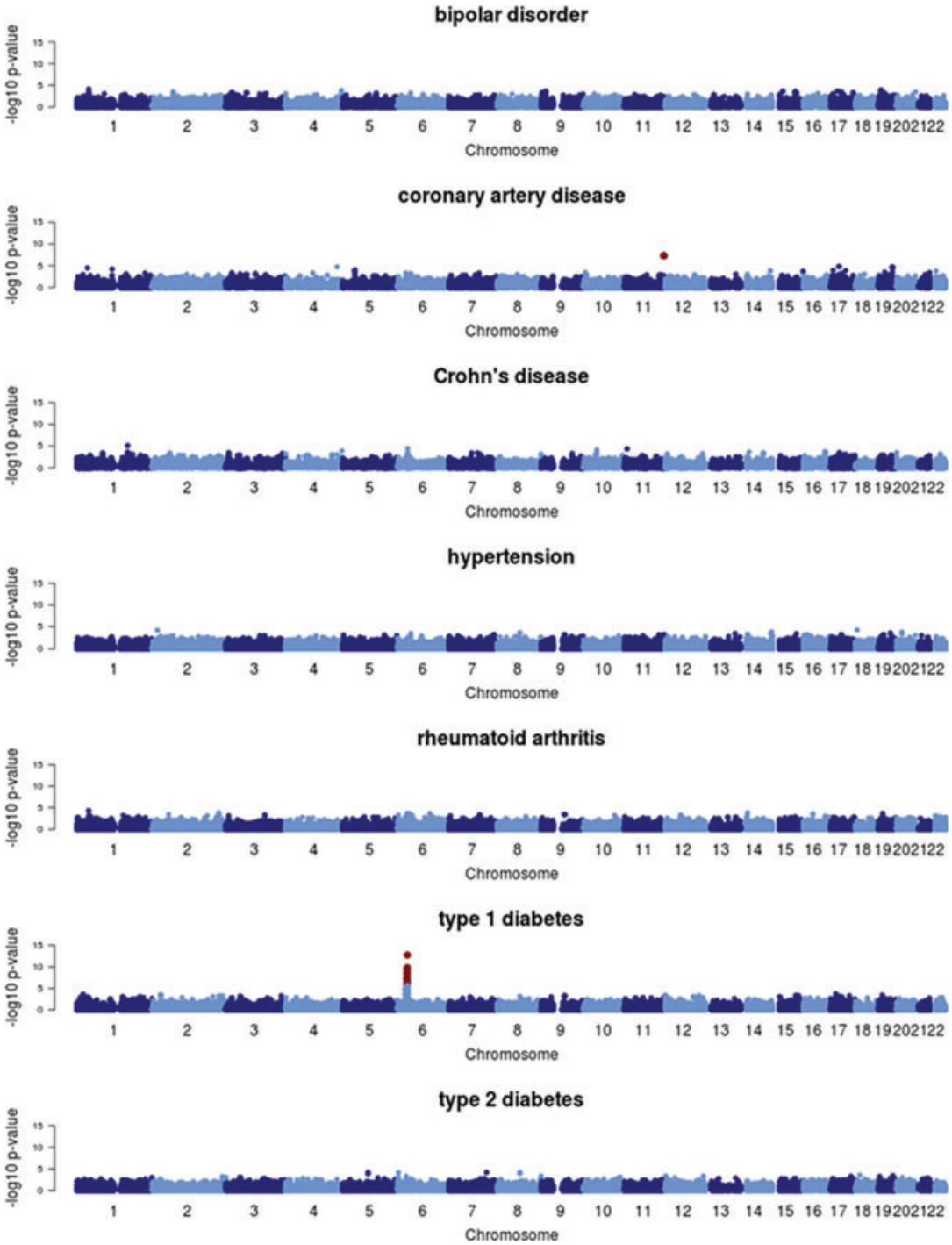

Fig. 3. Manhattan plots summarising association of seven diseases from the WTCCC experiment with accumulations of well-imputed rare variants (MAF $< 1\%$ and info score of at least 0.4) within genes (as defined by the UCSC human genome database). Each point represents a gene, plotted according to the observed $-\log_{10}$ *P*-value of association (*y*-axis) and the physical position of the midpoint of the transcript (*x*-axis), with those achieving genome-wide significance ($P < 1.7 \times 10^{-6}$) highlighted in red.

is adequately powered to detect. It has thus been suggested that much of the 'missing heritability' of complex human traits can be attributed to rare genetic variation [Bansal et al., 2010], which is not well captured by GWAS genotyping products. The most comprehensive approach to assaying rare genetic variation is through large-scale next-generation re-sequencing experiments. However, despite advances in the cost-effectiveness of these technologies, whole-genome re-sequencing of the large cohorts of individuals required to detect rare variant association with complex traits,

TABLE I. Genes demonstrating genome-wide significant ($P < 1.7 \times 10^{-6}$) evidence of rare variant association with type 1 diabetes in the MHC, before and after adjustment for the lead common GWAS SNP (rs9268645) in the region

Gene symbol	Build 37 chromosome 6 position (bp)		Number of rare variants	Mean MAF (%)	Analysis adjusted for three principal components only		Analysis adjusted for three principal components and rs9268645	
	Start	Stop			*P*-value	OR (95% CI)	*P*-value	OR (95% CI)
Genome-wide significant before adjustment for rs9268645								
HLA-DRA	32,407,646	32,412,821	23	0.32	2.0×10^{-13}	0.556 (0.476–0.650)	2.2×10^{-9}	0.642 (0.555–0.742)
HLA-DRB5	32,485,162	32,498,006	43	0.51	1.6×10^{-10}	0.746 (0.682–0.817)	1.6×10^{-10}	0.738 (0.673–0.810)
SLC44A4	31,837,321	31,846,823	27	0.22	1.7×10^{-10}	0.556 (0.465–0.666)	8.7×10^{-9}	0.586 (0.489–0.703)
PBX2	32,152,509	32,157,963	13	0.22	1.2×10^{-9}	0.375 (0.273–0.514)	9.5×10^{-14}	0.290 (0.210–0.402)
TNXA	31,976,196	31,981,050	7	0.41	2.6×10^{-9}	2.346 (1.772–3.107)	2.4×10^{-4}	1.719 (1.287–2.295)
AGPAT1	32,135,989	32,139,282	5	0.24	3.3×10^{-9}	0.118 (0.058–0.239)	7.2×10^{-7}	0.169 (0.084–0.342)
EHMT2	31,847,536	31,853,019	16	0.23	4.1×10^{-9}	0.437 (0.332–0.576)	5.1×10^{-7}	0.484 (0.365–0.643)
PBMUCL2	31,021,983	31,027,653	25	0.35	3.2×10^{-8}	0.757 (0.686–0.836)	9.1×10^{-7}	0.777 (0.702–0.859)
C6orf10	32,256,302	32,261,812	22	0.41	7.1×10^{-8}	0.748 (0.673–0.831)	1.9×10^{-5}	0.793 (0.713–0.882)
NCR3	31,557,050	31,560,762	10	0.23	1.0×10^{-6}	0.436 (0.312–0.608)	1.6×10^{-4}	0.518 (0.368–0.729)
Genome-wide significant after adjustment for rs9268645								
HLA-DMA	32,917,411	32,920,899	13	0.35	8.6×10^{-6}	0.606 (0.389–0.942)	1.1×10^{-7}	0.540 (0.430–0.678)
SKIV2L	31,926,580	31,937,532	33	0.20	4.0×10^{-5}	0.706 (0.507–0.984)	2.6×10^{-7}	0.640 (0.540–0.759)
TNXB	32,008,931	32,014,384	10	0.35	4.3×10^{-5}	0.589 (0.355–0.978)	4.1×10^{-7}	0.513 (0.396–0.664)

genome-wide, still represents an infeasible financial investment for most research groups.

We demonstrate here, by simulation, that imputation from an existing scaffold of GWAS genotype data using publicly available high-density reference panels, such as those made available through the 1000 Genomes Project [The 1000 Genomes Project Consortium, 2010], has the potential to identify rare variant associations with complex traits, without the need for costly re-sequencing experiments. These results are entirely consistent with other published simulation studies investigating the performance of rare variant association methodology using imputation up to re-sequencing data, either from an external reference panel [Li et al., 2010], or from a subset of the analysis cohort [Zawistowski et al., 2010]. Overall, our results suggest that a reference panel of 4,000 individuals offers noticeable gains in power over 500 individuals only when the spectrum of causal variants is loaded with rarer variants (i.e. in our simulations, when the maximum MAF of any individual causal variant is 0.5%, rather than 1%). In this scenario, imputation of the analysis cohort from the GWAS scaffold can achieve much of the power to detect rare variant association obtained by the gold-standard re-sequencing strategy. Our simulations assumed a GWAS scaffold with the same characteristics, in terms of allele frequency profile and density, as the Illumina Human660W-Quad BeadArray. We would expect the quality of imputed rare variants to be improved with more dense GWAS scaffolds, such as the Illumina HumanOmni5-Quad, although this evaluation is beyond the scope of this study.

We have considered a relatively simple underlying model for the association of the trait with multiple rare variants within the gene. More complex models might incorporate selection and/or different directions of effect of the causal variants on the trait. However, at present, we do not fully appreciate the likely effect of rare variants within a gene or pathway on complex human traits, although it is clear that the true underlying association model will impact the power of GRANVIL. Nevertheless, it is less obvious that the underlying association model will impact the relative performance of GRANVIL applied to rare variation derived from imputation as compared to that assayed through re-sequencing.

Our simulation study assumes that the analysis cohort and reference panel are ascertained from the same population, and thus are perfectly matched in terms of their rare variant profile. However, with publicly available reference panels, this is unlikely to be the case. Indeed, the Phase I 1000 Genomes Project reference panel (June 2011 interim release) consists of phased individuals from multiple populations that together incorporate a wide range of ancestry groups [The 1000 Genomes Project Consortium, 2010]. One cost-efficient approach to address this issue is to consider re-sequencing a small number of individuals from the analysis cohort to supplement the reference panel. This strategy has been successfully applied in identifying association of a population-specific imputed rare variant with sick sinus syndrome in Iceland [Holm et al., 2011].

Our simulation study also assumes that all rare causal variants in the gene have the same impact on the trait, for example, that they all result in loss-of-function of the gene product. The GRANVIL software makes the same underlying assumption, and hence power to detect rare variant association will be maximised. Under alternative models of rare variant association with the trait, which consider causal

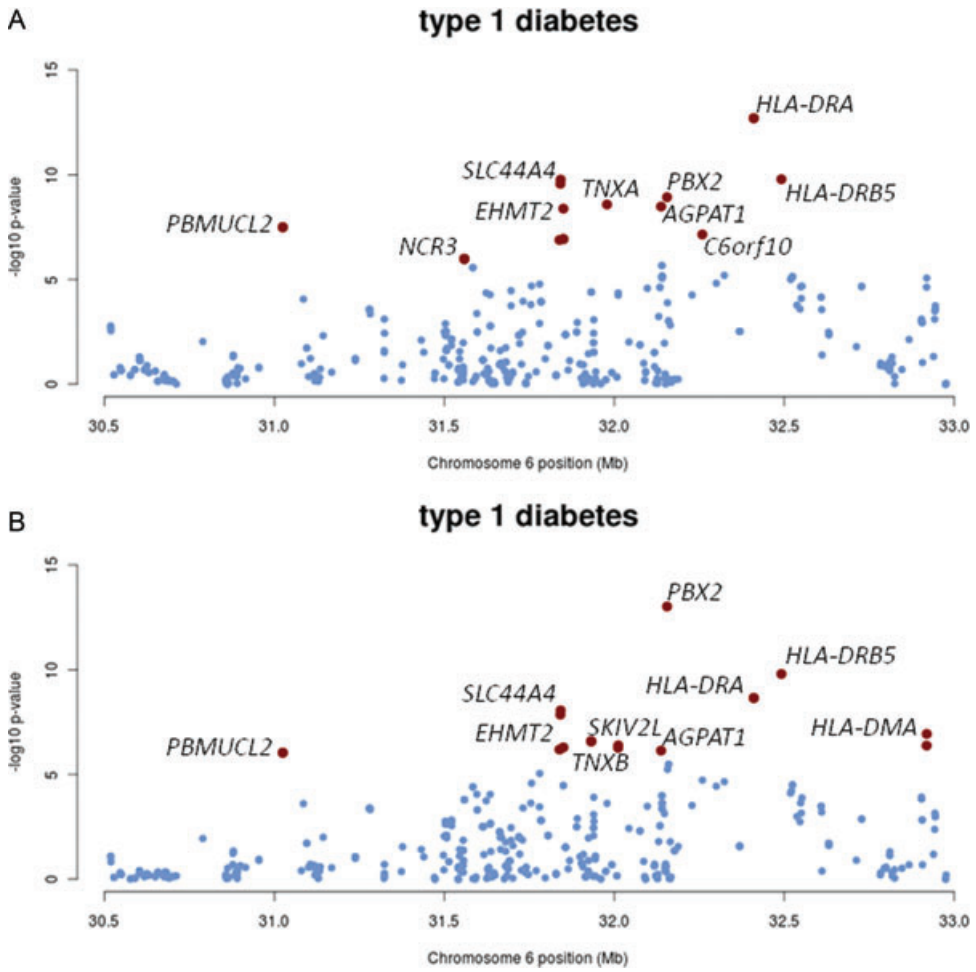

Fig. 4. Regional plots summarising association of type 1 diabetes with accumulations of well-imputed rare variants (MAF $< 1\%$ and info score of at least 0.4) within MHC genes (as defined by the UCSC human genome database). Each point represents a gene, and those achieving genome-wide significance ($P < 1.7 \times 10^{-6}$) are highlighted in red. The panels correspond to analyses (A) before and (B) after adjustment for the lead common GWAS SNP (rs9268645) in the region.

variants within the gene to result in both loss- and gain-of-function of the gene product, the power of GRANVIL will be reduced for all design strategies. For these models, powerful methods exist for detecting association with rare genetic variation [Asimit et al., 2012; Neale et al., 2011; Wu et al., 2011]. Many of these methods, such as C-alpha [Neale et al., 2011], require direct genotype calls. These approaches could make use of 'best guess' genotypes from imputation, although further development is required to appropriately allow for the posterior distribution of calls for imputed variants.

The encouraging results of our simulation study prompted us to re-assess the evidence of rare variant association with seven diseases from the WTCCC experiment [The Wellcome Trust Case Control Consortium, 2007]. We were able to recover genotypes at more than 5 million 'high-quality' imputed rare variants, even with the Affymetrix GeneChip 500K Mapping Array Set as a scaffold, which would not be expected to capture variation as well as more recent higher density genotyping products. Principal components analysis identified three axes of genetic variation that capture fine-scale population structure within the United Kingdom. After adjustment for these three components as covariates in our logistic regression modelling, there was no discernable residual inflation in rare variant association statistics, indicating that any additional fine-scale population structure had no impact on our analysis. We identified association of coronary artery disease with accumulations of minor alleles at rare variants in *PRDM10* at genome-wide significance. This gene has not been previously implicated in susceptibility to coronary artery disease or related cardio-metabolic phenotypes, and this association signals warrant follow-up in independent cohorts, either through genotyping of rare variants in the gene or re-sequencing. We also identified genome-wide significant evidence of association of type 1 diabetes with accumulations of minor alleles at rare variants in multiple genes from the MHC. This region has been previously associated with type 1 diabetes, both through common variant GWAS and analysis of classical human leukocyte antigen (HLA) haplotypes. Further work is required to dissect the complex genetic contribution of common and rare variation in this region to susceptibility to type 1 diabetes and other autoimmune disorders.

The results of our analyses presented here have major implications for the design and analysis of genome-wide rare variant association studies of complex human traits. Our results clearly highlight the potential for the detection of rare variant associations by using existing GWAS genotype data, supplemented with imputation from publicly available high-density reference panels, without the need for costly whole-genome re-sequencing experiments. Although imputation can never replace the gold-standard approach of whole-genome re-sequencing, it provides a powerful, cost-effective alternative that only requires a scaffold of GWAS genotype data, which may already be available. This message will bring encouragement to research groups who do not have sufficient funding to consider whole-genome, or even whole-exome sequencing as a financially viable approach to assaying rare genetic variation. It is clear that GWAS still have the potential to offer an exciting opportunity for gene discovery, conceivably leading to substantial advancements in our understanding of the genetic architecture underlying complex human traits.

ACKNOWLEDGMENTS

R.M. acknowledges support from the Development Fund of the University of Tartu, in the frame of the Centre of Transitional Genomics (grant SP1GVARENG). J.L.A., A.G.D.W., E.Z., and A.P.M. acknowledge support from the Wellcome Trust under awards WT098051, WT081682, and WT090532. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk. Funding for the Wellcome Trust Case Control Consortium project was provided by the Wellcome Trust under award WT076113. We acknowledge the Oxford Supercomputing Centre for providing computing facilities for data analysis.

REFERENCES

Ahituv N, Kavaslar N, Schackwitz W, Ustaszewska A, Martin J, Hebert S, Doelle H, Ersoy B, Kryukov G, Schmidt S, Yosef N, Ruppin E, Sharan R, Vaisse C, Sunyaev S, Dent R, Cohen J, McPherson R, Pennacchio LA. 2007. Medical sequencing at the extremes of human body mass. Am J Hum Genet 80:779–791.

Anderson CA, Pettersson FH, Barrett JC, Zhuang JJ, Ragoussis J, Cardon LR, Morris AP. 2008. Evaluating the effects of imputation on the power, coverage and cost-efficiency of genome-wide SNP platforms. Am J Hum Genet 83:112–119.

Asimit JL, Day-Williams AG, Morris AP, Zeggini E. 2012. ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data. Hum Hered 73:84–94.

Bansal V, Libiger O, Torkamani A, Schork NJ. 2010. Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet 11:773–785.

Barrett JC, Cardon LR. 2006. Evaluating coverage of genome-wide association studies. Nat Genet 38:659–662.

Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, Julier C, Morahan G, Nerup J, Nierras C, Plagnol V, Pociot F, Schuilenburg H, Smyth DJ, Stevens H, Todd JA, Walker NM, Rich SS; Type 1 Diabetes Genetics Consortium. 2009. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet 41:703–707.

Bodmer W, Bonilla C. 2008. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 40:695–701.

Browning BL, Browning SR. 2009. A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 81:1084–1097.

Cohen JC, Pertsemlidis A, Fahmi S, Esmail S, Vega GL, Grundy SM, Hobbs HH. 2006. Multiple rare variants in *NPC1L1* associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proc Natl Acad Sci USA 103:1810–1815.

Devlin B, Roeder K. 1999. Genomic control for association studies. Biometrics 55:997–1004.

Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. 2010. Rare variants create synthetic genome-wide associations. PLoS Biol 26: e1000294.

Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R, Anderson CA, Bis JC, Bumpstead S, Ellinghaus D, Festen EM, Georges M, Green T, Haritunians T, Jostins L, Latiano A, Mathew CG, Montgomery GW, Prescott NJ, Raychaudhuri S, Rotter JI, Schumm P, Sharma Y, Simms LA, Taylor KD, Whiteman D, Wijmenga C, Baldassano RN, Barclay M, Bayless TM, Brand S, Büning C, Cohen A, Colombel JF, Cottone M, Stronati L, Denson T, De Vos M, D'Inca R, Dubinsky M, Edwards C, Florin T, Franchimont D, Gearry R, Glas J, Van Gossum A, Guthery SL, Halfvarson J, Verspaget HW, Hugot JP, Karban A, Laukens D, Lawrance I, Lemann M, Levine A, Libioulle C, Louis E, Mowat C, Newman W, Panés J, Phillips A, Proctor DD, Regueiro M, Russell R, Rutgeerts P, Sanderson J, Sans M, Seibold F, Steinhart AH, Stokkers PC, Torkvist L, Kullak-Ublick G, Wilson D, Walters T, Targan SR, Brant SR, Rioux JD, D'Amato M, Weersma RK, Kugathasan S, Griffiths AM, Mansfield JC, Vermeire S, Duerr RH, Silverberg MS, Satsangi J, Schreiber S, Cho JH, Annese V, Hakonarson H, Daly MJ, Parkes M. 2010. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet 42:1118–1125.

Frazer KA, Murray SS, Schork NJ, Topol EJ. 2009. Human genetic variation and its contribution to complex traits. Nat Rev Genet 10:241–251.

Griffiths RD, Marjoram P. 2007. An Ancestral Recombination Graph. Progress in Population Genetics and Human Evolution. New York: Springer-Verlag, p. 257–270.

Han F, Pan W. 2010. A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered 70:42–54.

Hao K, Li C, Rosenow C, Hung Wong W. 2004. Estimation of genotype error rate using samples with pedigree information – an application on the GeneChip Mapping 10K array. Genomics 84:623–630.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106:9362–9367.

Hoffman TJ, Marini NJ, Witte JS. 2010. Comprehensive approach to analyzing rare genetic variants. PLoS One 5:e13584.

Holm H, Gudbjartsson DF, Sulem P, Masson G, Helgadottir HT, Zanon C, Magnusson OT, Helgason A, Saemundsdottir J, Gylfason A, Stefansdottir H, Gretarsdottir S, Matthiasson SE, Thorgeirsson GM, Jonasdottir A, Sigurdsson A, Stefansson H, Werge T, Rafnar T, Kiemeney LA, Parvez B, Muhammad R, Roden DM, Darbar D, Thorleifsson G, Walters GB, Kong A, Thorsteinsdottir U, Arnar DO, Stefansson K. 2011. A rare variant in *MYH6* is associated with high risk of sick sinus syndrome. Nat Genet 43:316–320.

Howie BN, Donnelly P, Marchini J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5:e1000529.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model. Bioinformatics 18:337–338.

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861.

Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, Simon DB, Newton–Cheh C, State MW, Levy D, Lifton RP. 2008. Rare independent mutations in renal salt genes contribute to blood pressure variation. Nat Genet 40:592–599.

Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, Ferreira T, Wood AR, Weyant RJ, Segrè AV, Speliotes EK, Wheeler E, Soranzo N, Park JH, Yang J, Gudbjartsson D, Heard-Costa NL, Randall JC, Qi L, Vernon Smith A, Mägi R, Pastinen T, Liang L, Heid IM, Luan J, Thorleifsson G, Winkler TW, Goddard ME, Sin Lo K, Palmer C, Workalemahu T, Aulchenko YS, Johansson A, Zillikens MC, Feitosa MF, Esko T, Johnson T, Ketkar S, Kraft P, Mangino M, Prokopenko I, Absher D, Albrecht E, Ernst F, Glazer NL, Hayward C, Hottenga JJ, Jacobs KB, Knowles JW, Kutalik Z, Monda KL, Polasek O, Preuss M, Rayner NW, Robertson NR, Steinthorsdottir V, Tyrer JP, Voight BF, Wiklund F, Xu J, Zhao JH, Nyholt DR, Pellikka N, Perola M, Perry JR, Surakka I, Tammesoo ML, Altmaier EL, Amin N, Aspelund T, Bhangale T, Boucher G, Chasman DI, Chen C, Coin L, Cooper MN, Dixon AL, Gibson Q, Grundberg E, Hao K, Juhani Junttila M, Kaplan LM, Kettunen J, König IR, Kwan T, Lawrence RW, Levinson DF, Lorentzon M, McKnight B, Morris AP, Müller M, Suh Ngwa J, Purcell S, Rafelt S, Salem RM, Salvi E, Sanna S, Shi J, Sovio U, Thompson JR, Turchin MC, Vandenput L, Verlaan DJ, Vitart V, White CC, Ziegler A, Almgren P, Balmforth AJ, Campbell H, Citterio L, De Grandi A, Dominiczak A, Duan J, Elliott P, Elosua R, Eriksson JG, Freimer NB, Geus EJ, Glorioso N, Haiqing S, Hartikainen AL, Havulinna AS, Hicks AA, Hui J, Igl W, Illig T, Jula A, Kajantie E, Kilpeläinen TO, Koiranen M, Kolcic I, Koskinen S, Kovacs P, Laitinen J, Liu J, Lokki ML, Marusic A, Maschio A, Meitinger T, Mulas A, Paré G, Parker AN, Peden JF, Petersmann A, Pichler I, Pietiläinen KH, Pouta A, Ridderstråle M, Rotter JI, Sambrook JG, Sanders AR, Schmidt CO, Sinisalo J, Smit JH, Stringham HM, Bragi Walters G, Widen E, Wild SH, Willemsen G, Zagato L, Zgaga L, Zitting P, Alavere H, Farrall M, McArdle WL, Nelis M, Peters MJ, Ripatti S, van Meurs JB, Aben KK, Ardlie KG, Beckmann JS, Beilby JP, Bergman RN, Bergmann S, Collins FS, Cusi D, den Heijer M, Eiriksdottir G, Gejman PV, Hall AS, Hamsten A, Huikuri HV, Iribarren C, Kähönen M, Kaprio J, Kathiresan S, Kiemeney L, Kocher T, Launer LJ, Lehtimäki T, Melander O, Mosley TH Jr, Musk AW, Nieminen MS, O'Donnell CJ, Ohlsson C, Oostra B, Palmer LJ, Raitakari O, Ridker PM, Rioux JD, Rissanen A, Rivolta C, Schunkert H, Shuldiner AR, Siscovick DS, Stumvoll M, Tönjes A, Tuomilehto J, van Ommen GJ, Viikari J, Heath AC, Martin NG, Montgomery GW, Province MA, Kayser M, Arnold AM, Atwood LD, Boerwinkle E, Chanock SJ, Deloukas P, Gieger C, Grönberg H, Hall P, Hattersley AT, Hengstenberg C, Hoffman W, Lathrop GM, Salomaa V, Schreiber S, Uda M, Waterworth D, Wright AF, Assimes TL, Barroso I, Hofman A, Mohlke KL, Boomsma DI, Caulfield MJ, Cupples LA, Erdmann J, Fox CS, Gudnason V, Gyllensten U, Harris TB, Hayes RB, Jarvelin MR, Mooser V, Munroe PB, Ouwehand WH, Penninx BW, Pramstaller PP, Quertermous T, Rudan I, Samani NJ, Spector TD, Völzke H, Watkins H, Wilson JF, Groop LC, Haritunians T, Hu FB, Kaplan RC, Metspalu A, North KE, Schlessinger D, Wareham NJ, Hunter DJ, O'Connell JR, Strachan DP, Wichmann HE, Borecki IB, van Duijn CM, Schadt EE, Thorsteinsdottir U, Peltonen L, Uitterlinden AG, Visscher PM, Chatterjee N, Loos RJ, Boehnke M, McCarthy MI, Ingelsson E, Lindgren CM, Abecasis GR, Stefansson K, Frayling TM, Hirschhorn JN. 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467:832–838.

Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83:311–321.

Li Y, Byrnes AE, Li M. 2010. To identify associations with rare variants, just WHaIT: weighted haplotype and imputation-based tests. Am J Hum Genet 87:728–735.

Madsen BE, Browning SR. 2009. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5: e1000384.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. 2009. Finding the missing heritability of complex diseases. Nature 461:747–753.

Marchini J, Howie B. 2010. Genotype imputation for genome-wide association studies. Nat Rev Genet 11:499–511.

Metzker ML. 2010. Sequencing technologies: the next generation. Nat Rev Genet 11:31–46.

Morgenthaler S, Thilly WG. 2007. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res 615:28–56.

Morris AP, Zeggini E. 2010. An evaluation of statistical approaches to rare variant association analysis in genetic association studies. Genet Epidemiol 34:188–195.

Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. 2011. Testing for an unusual distribution of rare variants. PLoS Genet 7: e1001322.

Nejentsev S, Howson JM, Walker NM, Szeszko J, Field SF, Stevens HE, Reynolds P, Hardy M, King E, Masters J, Hulme J, Maier LM, Smyth D, Bailey R, Cooper JD, Ribas G, Campbell RD, Clayton DG, Todd JA; Wellcome Trust Case Control Consortium. 2007. Localisation of type 1 diabetes susceptibility to the MHC class I genes *HLA-B* and *HLA-A*. Nature 450:887–892.

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ. 2010. Exome sequencing identifies the cause of a Mendelian disorder. Nat Genet 42:30–35.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909.

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. 2010. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86:832–838.

Pritchard JK. 2001. Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69:124–137.

Romeo S, Pennacchio LA, Fu Y, Boerwinkle E, Tybjaerg-Hansen A, Hobbs HH, Cohen JC. 2007. Population-based resequencing of *ANGPTL4* uncovers variation that reduce triglycerides and increase HDL. Nat Genet 39:513–516.

Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Allen HL, Lindgren CM, Luan J, Mägi R, Randall JC, Vedantam S, Winkler TW, Qi L, Workalemahu T, Heid IM, Steinthorsdottir V, Stringham HM, Weedon MN, Wheeler E, Wood AR, Ferreira T, Weyant RJ, Segrè AV, Estrada K, Liang L, Nemesh J, Park JH, Gustafsson S, Kilpeläinen TO, Yang J, Bouatia-Naji N, Esko T, Feitosa MF, Kutalik Z, Mangino M, Raychaudhuri S, Scherag A, Smith AV, Welch R, Zhao JH, Aben KK, Absher DM, Amin N, Dixon AL, Fisher E, Glazer NL, Goddard ME, Heard-Costa NL, Hoesel V, Hottenga JJ, Johansson A, Johnson T, Ketkar S, Lamina C, Li S, Moffatt MF, Myers RH, Narisu N, Perry JR, Peters MJ, Preuss M, Ripatti S, Rivadeneira F, Sandholt C, Scott LJ, Timpson NJ, Tyrer JP, van Wingerden S, Watanabe RM, White CC, Wiklund F, Barlassina C, Chasman DI, Cooper MN, Jansson JO, Lawrence RW, Pellikka N, Prokopenko I, Shi J, Thiering E, Alavere H, Alibrandi MT, Almgren P, Arnold AM, Aspelund T, Atwood LD, Balkau B, Balmforth AJ, Bennett AJ, Ben-Shlomo Y, Bergman RN, Bergmann S, Biebermann H, Blakemore AI, Boes T, Bonnycastle LL, Bornstein SR, Brown MJ, Buchanan TA, Busonero F, Campbell H, Cappuccio FP, Cavalcanti-Proença C, Chen YD, Chen CM, Chines PS, Clarke R, Coin L, Connell J, Day IN, den Heijer M, Duan J, Ebrahim S, Elliott P, Elosua R, Eiriksdottir G, Erdos MR, Eriksson JG, Facheris MF, Felix SB, Fischer-Posovszky P, Folsom AR, Friedrich N, Freimer NB, Fu M, Gaget S, Gejman PV, Geus EJ, Gieger C, Gjesing AP, Goel A, Goyette P, Grallert H, Grässler J, Greenawalt DM, Groves CJ, Gudnason V, Guiducci C, Hartikainen AL, Hassanali N, Hall AS, Havulinna AS, Hayward C, Heath AC, Hengstenberg C, Hicks AA, Hinney A, Hofman A, Homuth G, Hui J, Igl W, Iribarren C, Isomaa B, Jacobs KB, Jarick I, Jewell E, John U, Jørgensen T, Jousilahti P, Jula A, Kaakinen M, Kajantie E, Kaplan LM,

Kathiresan S, Kettunen J, Kinnunen L, Knowles JW, Kolcic I, König IR, Koskinen S, Kovacs P, Kuusisto J, Kraft P, Kvaløy K, Laitinen J, Lantieri O, Lanzani C, Launer LJ, Lecoeur C, Lehtimäki T, Lettre G, Liu J, Lokki ML, Lorentzon M, Luben RN, Ludwig B; MAGIC, Manunta P, Marek D, Marre M, Martin NG, McArdle WL, McCarthy A, McKnight B, Meitinger T, Melander O, Meyre D, Midthjell K, Montgomery GW, Morken MA, Morris AP, Mulic R, Ngwa JS, Nelis M, Neville MJ, Nyholt DR, O'Donnell CJ, O'Rahilly S, Ong KK, Oostra B, Paré G, Parker AN, Perola M, Pichler I, Pietiläinen KH, Platou CG, Polasek O, Pouta A, Rafelt S, Raitakari O, Rayner NW, Ridderstråle M, Rief W, Ruokonen A, Robertson NR, Rzehak P, Salomaa V, Sanders AR, Sandhu MS, Sanna S, Saramies J, Savolainen MJ, Scherag S, Schipf S, Schreiber S, Schunkert H, Silander K, Sinisalo J, Siscovick DS, Smit JH, Soranzo N, Sovio U, Stephens J, Surakka I, Swift AJ, Tammesoo ML, Tardif JC, Teder-Laving M, Teslovich TM, Thompson JR, Thomson B, Tönjes A, Tuomi T, van Meurs JB, van Ommen GJ, Vatin V, Viikari J, Visvikis-Siest S, Vitart V, Vogel CI, Voight BF, Waite LL, Wallaschofski H, Walters GB, Widen E, Wiegand S, Wild SH, Willemsen G, Witte DR, Witteman JC, Xu J, Zhang Q, Zgaga L, Ziegler A, Zitting P, Beilby JP, Farooqi IS, Hebebrand J, Huikuri HV, James AL, Kähönen M, Levinson DF, Macciardi F, Nieminen MS, Ohlsson C, Palmer LJ, Ridker PM, Stumvoll M, Beckmann JS, Boeing H, Boerwinkle E, Boomsma DI, Caulfield MJ, Chanock SJ, Collins FS, Cupples LA, Smith GD, Erdmann J, Froguel P, Grönberg H, Gyllensten U, Hall P, Hansen T, Harris TB, Hattersley AT, Hayes RB, Heinrich J, Hu FB, Hveem K, Illig T, Jarvelin MR, Kaprio J, Karpe F, Khaw KT, Kiemeney LA, Krude H, Laakso M, Lawlor DA, Metspalu A, Munroe PB, Ouwehand WH, Pedersen O, Penninx BW, Peters A, Pramstaller PP, Quertermous T, Reinehr T, Rissanen A, Rudan I, Samani NJ, Schwarz PE, Shuldiner AR, Spector TD, Tuomilehto J, Uda M, Uitterlinden A, Valle TT, Wabitsch M, Waeber G, Wareham NJ, Watkins H; Procardis Consortium, Wilson JF, Wright AF, Zillikens MC, Chatterjee N, McCarroll SA, Purcell S, Schadt EE, Visscher PM, Assimes TL, Borecki IB, Deloukas P, Fox CS, Groop LC, Haritunians T, Hunter DJ, Kaplan RC, Mohlke KL, O'Connell JR, Peltonen L, Schlessinger D, Strachan DP, van Duijn CM, Wichmann HE, Frayling TM, Thorsteinsdottir U, Abecasis GR, Barroso I, Boehnke M, Stefansson K, North KE, McCarthy MI, Hirschhorn JN, Ingelsson E, Loos RJ. 2010. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet 42:937–948.

The 1000 Genomes Project Consortium. 2010. A map of human sequence variation from population-scale sequencing. Nature 467:1061–1073.

Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G, McCulloch LJ, Ferreira T, Grallert H, Amin N, Wu G, Willer CJ, Raychaudhuri S, McCarroll SA, Langenberg C, Hofmann OM, Dupuis J, Qi L, Segrè AV, van Hoek M, Navarro P, Ardlie K, Balkau B, Benediktsson R, Bennett AJ, Blagieva R, Boerwinkle E, Bonnycastle LL, Bengtsson Boström K, Bravenboer B, Bumpstead S, Burtt NP, Charpentier G, Chines PS, Cornelis M, Couper DJ, Crawford G, Doney AS, Elliott KS, Elliott AL, Erdos MR, Fox CS, Franklin CS, Ganser M, Gieger C, Grarup N, Green T, Griffin S, Groves CJ, Guiducci C, Hadjadj S, Hassanali N, Herder C, Isomaa B, Jackson AU, Johnson PR, Jørgensen T, Kao WH, Klopp N, Kong A, Kraft P, Kuusisto J, Lauritzen T, Li M, Lieverse A, Lindgren CM, Lyssenko V, Marre M, Meitinger T, Midthjell K, Morken MA, Narisu N, Nilsson P, Owen KR, Payne F, Perry JR, Petersen AK, Platou C, Proença C, Prokopenko I, Rathmann W, Rayner NW, Robertson NR, Rocheleau G, Roden M, Sampson MJ, Saxena R, Shields BM, Shrader P, Sigurdsson G, Sparsø T, Strassburger K, Stringham HM, Sun Q, Swift AJ, Thorand B, Tichet J, Tuomi T, van Dam RM, van Haeften TW, van Herpt T, van Vliet-Ostaptchouk JV, Walters GB, Weedon MN, Wijmenga C, Witteman J, Bergman RN, Cauchi S, Collins FS, Gloyn AL, Gyllensten U, Hansen T, Hide WA, Hitman GA, Hofman A, Hunter DJ, Hveem K, Laakso M, Mohlke KL, Morris AD, Palmer CN, Pramstaller PP, Rudan I, Sijbrands E, Stein LD, Tuomilehto J, Uitterlinden A, Walker M, Wareham NJ, Watanabe RM, Abecasis GR, Boehm BO, Campbell H, Daly MJ, Hattersley AT, Hu FB, Meigs JB, Pankow JS, Pedersen O, Wichmann HE, Barroso I, Florez JC, Frayling TM, Groop L, Sladek R, Thorsteinsdottir U, Wilson JF, Illig T, Froguel P, van Duijn CM, Stefansson K, Altshuler D, Boehnke M, McCarthy MI; MAGIC investigators; GIANT Consortium. 2010. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet 42:579–589.

The Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequence data with the sequence kernel association test. Am J Hum Genet 89:82–93.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM. 2010. Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42:565–569.

Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zöllner S. 2010. Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. Am J Hum Genet 87:604–617.

APPENDIX: CALCULATION OF THE RESIDUAL TRAIT VARIANCE

Determine the number of minor alleles carried by the *i*th individual in the analysis cohort across all causal variants in the 50 kb gene, denoted m_i. The expected trait value for the *i*th individual, denoted by μ_i, takes the value 1 if $m_i > 0$, and 0 otherwise. Assuming an analysis cohort of N individuals, the mean expected trait value is given by

$$M = \sum_{i=1}^{N} \frac{\mu_i}{N}$$

with corresponding genetic variance given by

$$V_G = \sum_{i=1}^{N} \frac{(\mu_i - M)^2}{N}.$$

Assuming that the rare causal variants explain 100λ% of the overall trait variance, it follows that the residual variance is given by

$$\sigma^2 = V_G \frac{(1-\lambda)}{\lambda}.$$