European Journal of Human Genetics (2012), 1–5

www.nature.com/ejhg

SHORT REPORT

1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data

Jie Huang[1], David Ellinghaus[2], Andre Franke[2], Bryan Howie[3] and Yun Li*[,4]

We hypothesize that imputation based on data from the 1000 Genomes Project can identify novel association signals on a genome-wide scale due to the dense marker map and the large number of haplotypes. To test the hypothesis, the Wellcome Trust Case Control Consortium (WTCCC) Phase I genotype data were imputed using 1000 genomes as reference (20100804 EUR), and seven case/control association studies were performed using imputed dosages. We observed two 'missed' disease-associated variants that were undetectable by the original WTCCC analysis, but were reported by later studies after the 2007 WTCCC publication. One is within the *IL2RA* gene for association with type 1 diabetes and the other in proximity with the *CDKN2B* gene for association with type 2 diabetes. We also identified two refined associations. One is SNP rs11209026 in exon 9 of *IL23R* for association with Crohn's disease, which is predicted to be probably damaging by PolyPhen2. The other refined variant is in the *CUX2* gene region for association with type 1 diabetes, where the newly identified top SNP rs1265564 has an association *P*-value of 1.68×10^{-16}. The new lead SNP for the two refined loci provides a more plausible explanation for the disease association. We demonstrated that 1000 Genomes-based imputation could indeed identify both novel (in our case, 'missed' because they were detected and replicated by studies after 2007) and refined signals. We anticipate the findings derived from this study to provide timely information when individual groups and consortia are beginning to engage in 1000 genomes-based imputation.

European Journal of Human Genetics advance online publication, 1 February 2012; doi:10.1038/ejhg.2012.3

Keywords: genome-wide association study; the 1000 Genomes project; imputation

INTRODUCTION

It has been four years since the publication of one of the first and largest genome-wise association studies (GWAS), the Wellcome Trust Case Control Consortium (WTCCC) study.[1] Although imputation-based analysis was already adopted at that time for refining association signals, its use was limited and few results were reported based on imputations. In one study, an imputed missense mutation in gene *GCKR* is found to be more strongly associated with triglyceride.[2]

Imputation significantly increases the statistical power of GWAS and allows meta-analysis of studies genotyped on different platforms.[3–5] Until recently, most imputation work has been using HapMap haplotypes as a reference panel.[6] However, the recent publication of the 1000 Genomes Pilot Project and the availability of phased haplotypes from both the pilot and main project bring opportunities for much denser imputations and more extensive analysis.[7] The 1000 genomes-based imputation has already been shown to refine association signals and identify underlying genetic variants that are in high linkage disequilibrium (LD) with variants that were included in the genotyping platform.[7] However, there have been few reports of novel findings using this latest imputation approach on a genome-wide scale.[8,9] The Oxford-GSK study[8] used 1000 genomes-based imputation to refine a single genetic region and successfully identified a SNP that is in the promoter region of a biologically plausible gene with a more significant *P*-value than that without 1000 genomes imputation. The Sardiana study[9] fully utilized the reference panels from HapMap2, HapMap3, and the 1000 genomes, but did not specifically evaluate the power gains from each panel. It also has a smaller discovery sample size ($N<1700$) compared with the WTCCC study. In contrast, our study is hypothesis driven, based on a flagship study with rich phenotypes and a large sample size. We reason it is very important to confirm that genotype imputation based on the latest 1000 genomes release could identify novel variants on a genome-wide scale, beside refining associations at a regional level, given the large amount of efforts needed for scientists around the globe poised to apply this emerging tool to their scientific investigation.

We hypothesize that 1000 genomes-based imputation can identify novel variants beyond what could be seen from purely genotyped data or HapMap imputed data, due to the much denser SNP coverage and a much broader representation of reference populations. We test this hypothesis by re-examining the WTCCC Phase I data after imputing the genotype data to the full set of SNPs present in 1000 genomes latest release (version 20100804). We re-run association analysis for the seven traits based on 1000 genomes imputed dosages and highlight novel and refined genetic associations that would have been discovered by the original study should the 1000 genomes reference panel be available back then.

[1]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK; [2]Institute of Clinical Molecular Biology, Christian-Albrechts-University, Kiel, Germany; [3]Department of Human Genetics, University of Chicago, Chicago, IL, USA; [4]Department of Genetics, Department of Biostatistics, University of North Carolina, Chapel Hill, NC, USA
*Correspondence: Dr Y Li, Department of Genetics, Department of Biostatistics, University of North Carolina Chapel Hill, Chapel Hill, NC, 27599-7264, USA.
Tel: +919 843 2832; Fax: +919 843 4682; E-mail: yunli@med.unc.edu
Received 7 September 2011; revised 30 December 2011; accepted 4 January 2012

METHODS

We obtained approval for using the raw genotype data for the original WTCCC data set, and we created one harmonized genotype data set by applying quality controls similar as the original study to the downloaded files (details provided in Online Supplementary Section S1). Using this genotype data with embedded disease status, we ran the case/control association tests with PLINK and verified that the results are similar as those reported in the original study with negligible difference.[10] To match the genomic position used by the 1000 genomes reference, we mapped all SNPs to NCBI's build37 (hg19).

We used the MaCH program to first phase the haplotypes and then ran MiniMac for genotype imputation.[11] We used the recommended two-step approach and recommended parameters of 20 iterations of the Markov sampler and 200 states. The 1000 genomes reference panel was obtained from the University of Michigan Abecasis lab, version 20100804 (http://www.sph.umich.edu/csg/abecasis/MACH/download/1000G-2010-08.html). A total of 566 reference haplotypes of the European ancestry served as the reference panel. We ran a logistic regression analysis based on imputation dosages via MACH2DAT (thus taking imputation uncertainty into account) for each of the seven traits with the shared control without covariate adjustment, a similar statistical analysis approach presented by the original WTCCC study. We included all SNPs with estimated R^2 >0.3 and minor allele frequency >0.01 for analysis.

SNPs were considered in the same region if they are within the same gene or are <1 Mb apart. We define 'novel' for any SNP with association *P*-value reaching the genome-wide significance level (5×10^{-8}) in a region not reported in the original study that analyzed genotyped data. Novel variants that were later reported by other studies after the original 2007 WTCCC paper are designated as 'missed' instead of 'novel'. We define 'refined' for any association where there is a reported association in the same region in the original study but the new lead SNP has a *P*-value more significant even after correcting the number of new SNPs tested. For 'refined' association, we further require that the lead SNP has either a functional support or is pinpointing to biologically more relevant gene. We used PolyPhen-2 to predict the possible impact of amino-acid substitutions *in silico*.[12] To evaluate whether the genome-wide significant threshold of 5×10^{-8} widely used in HapMap2 imputed analysis would be sufficiently conservative in 1000 genomes imputed analysis, we picked tagging SNPs using a greedy algorithm similar to that in *ldselect*[13] at R^2=0.9 for the SNPs included in our analysis.

For all genetic loci identified as novel or refined, a 5-Mb region around the lead SNP was re-imputed and then analyzed with an independent imputation program (BEAGLE)[14] and association analysis tool (PLINK)[10] by an independent analyst (DE; Supplementary Section S2).

RESULTS

After SNP quality control and mapping of genomic positions to build37, a total of 389 827 SNPs for 16 179 samples were retained as input genotype data for imputation. After imputation, a total of 6 233 112 SNPs with estimated R^2 >0.3 and minor allele frequency >0.01 were used for association analysis. The estimated genomic inflation factor λ[15] for the seven case/control GWAS ranges from 1.04 to 1.09, which is comparable to the original study and indicates low genomic inflation. Association Manhattan plots for all seven analyses are shown in Supplementary Figure S1. We compared the signals with those in the original WTCCC study, and highlighted two missed and two refined variants that we identified through this latest imputation method (Table 1). All four loci were confirmed by an independent analysis using BEAGLE and PLINK (JH, DE).

As shown in the regional plots, the two missed variants would not have been identified with HapMap2-based[16] imputation (shown as red color in Figure 1). For the refined *CUX2* region, the best HapMap2 imputed SNP is less significant than the genotyped SNP originally reported. For the refined *IL23R* region, the best HapMap2 SNP is not exonic and not predicted to have functional consequence.

A total of 1 915 543 tagging SNPs were picked for the total of 6 233 112 SNPs included in our analysis, at the R^2 threshold of 0.9.

Table 1 Novel/missed and refined variants from 1000 genomes imputation-based analysis

Type	*Trait*	*Lead SNP in present study (association P[a])*	*Chromosome band*	*Gene*	*MAF*	*Estimated R^2[b]*	*Alleles*	*Odds ratio (95% CI)*	*Lead SNP in original Study (association P, LD[c], physical distance, conditional P[d])*	*Nearest SNP in literature (association P, LD[c], physical distance, conditional P[d])*
Missed	T1D	**rs61839660** $P=5.1\times10^{-9}$ (1.8×10^{-8})	10p15	*IL2RA* intron 7	0.09	0.87 (0.90)	C->T	1.60 (1.44–1.76)	NA	**rs12251307**[18] ($P=2.9\times10^{-5}$; LD=0.74 (0.38) 28.8 kb *Pcon*=3.7×10^{-5})
Missed	T2D	**rs7018475** $P=2.5\times10^{-8}$ (6.1×10^{-8})	9p21	*CDKN2B* -128 kb	0.24	0.79 (0.85)	T->G	0.74 (0.64–0.85)	NA	**rs10757282**[20] ($P=2.7\times10^{-3}$; LD=0.97 (0.41) 3.7 kb *Pcon*=3.7×10^{-7}) **rs10811661**[19,20] ($P=7.1\times10^{-4}$; LD=0.50 (0.02) 3.6 kb *Pcon*=1.5×10^{-7}) *Pcon*=0.5[e]
Refined	CD	**rs11209026** $P=4.2\times10^{-21}$ (1.1×10^{-17})	1p31	*IL23R* exon 9	0.06	0.90 (0.91)	G->A	3.18 (2.91–3.44)	**rs11805303** ($P=5.8\times10^{-12}$; LD=0.57 (0.01) 30.4 kb *Pcon*=7.6×10^{-17})	**rs11465804**[22] ($P=7.7\times10^{-20}$; LD=1.0 (0.88) 3.4 kb *Pcon*=7.9×10^{-3})
Refined	T1D	**rs1265564** $P=1.0\times10^{-16}$ (1.1×10^{-16})	12q24	*CUX2* intron 4	0.48	0.87 (0.92)	A->C	0.69 (0.60–0.78)	**rs17696736** ($P=1.5\times10^{-14}$; LD=0.43 (0.18) 778.4 kb *Pcon*=5.3×10^{-6})	**rs3184504**[18] ($P=1.4\times10^{-14}$; LD=0.74 (0.47) 176.2 kb *Pcon*=1.1×10^{-4})

Abbreviations: CD: Crohns's disease; T1D: type 1 diabetes; T2D: type 2 diabetes.
[a]The first value is based on MACH/MACH2DAT and the value in brackets is based on BEAGLE/PLINK.
[b]Imputation quality (estimated R^2) based on MaCH and Beagle (values in brackets).
[c]LD between lead SNP reported in this study and SNP in this column based on the same version of the 1000 genomes data, measured by D-prime (R^2).
[d]Conditional *P*-value of the lead SNP reported in this study conditioned on SNP in this column.
[e]Conditional *P*-value of the lead SNP on the two-SNP (rs10757282 and rs10811661) haplotype.

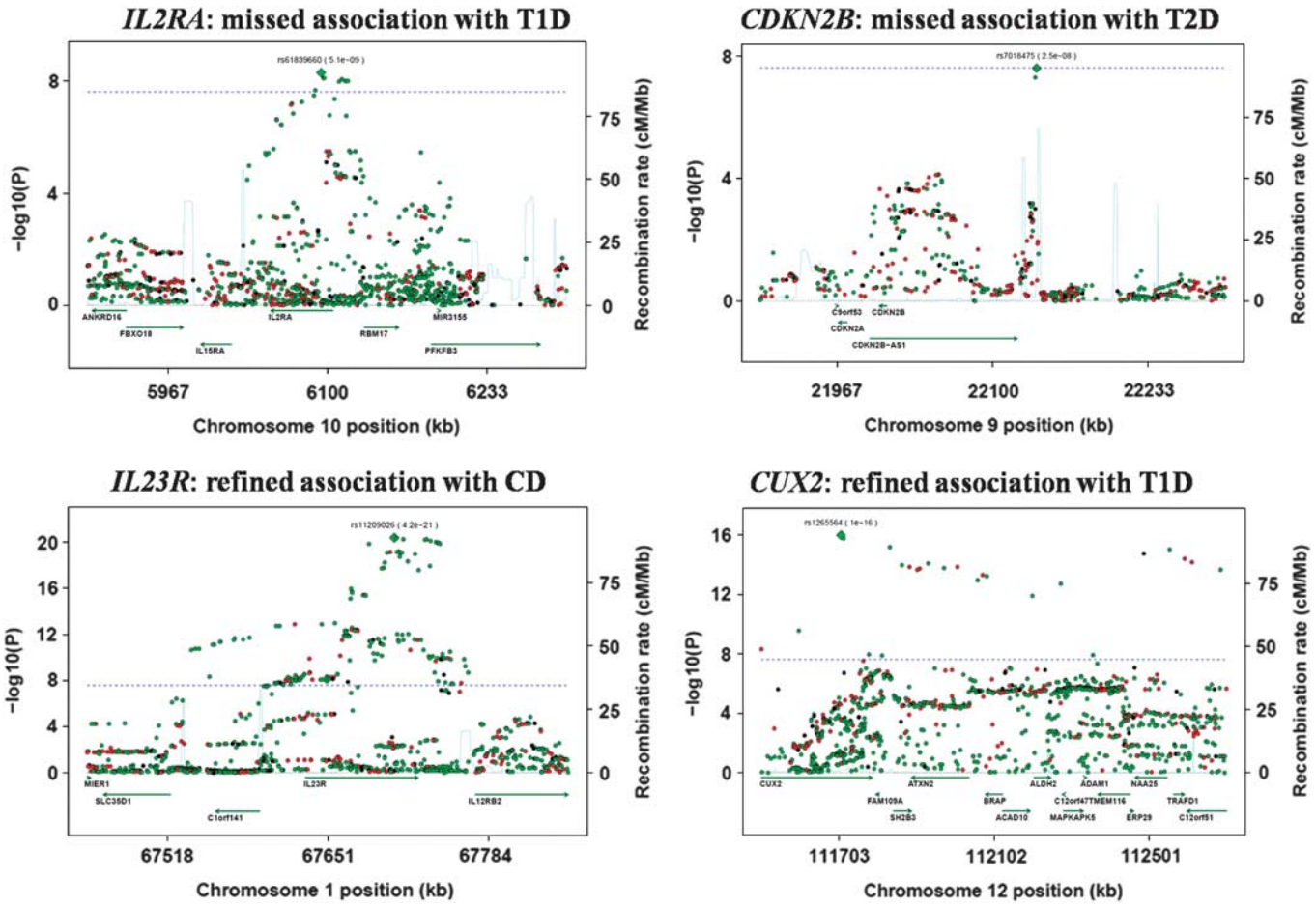

Figure 1 Regional plots for two novel (missed) and two refined loci. The top two plots are for two novel ('missed') regions, where highly significant SNPs meet genome-wide significance ($P<2.5\times10^{-8}$). The bottom two plots are for two refined regions. SNPs are represented by three different colors: black for WTCCC genotyped SNPs, red for HapMap2 imputed SNPs, and green for 1000 genomes imputed SNPs. Chromosome base pair positions (NCBI build 37) are represented on the *X*-axis. On the *Y*-axis, statistical significance is expressed as $-\log_{10}$ of the *P*-values. The horizontal line marks the $P=2.5\times10^{-8}$ threshold of genome-wide significance.

Therefore, compared with the previous assumption of one million independent loci across the genome,[17] and a genome-wide significance threshold of 5.0×10^{-8}, we propose a genome-wide threshold of 2.5×10^{-8} assuming two million independent SNPs. All four SNPs in Table 1 meet genome-wide significance applying this conservative (because of correlation among the tagging SNPs) threshold.

DISCUSSION

This is the first study to comprehensively assess the utility of 1000 genomes-based imputation for identifying novel genetic association signals. We identified two associations that were not reported in the original WTCCC study but later established through other GWAS. One is within gene *IL2RA* for association with type 1 diabetes (T1D) and the other is 128 kb downstream of gene *CDKN2B* for association with type 2 diabetes (T2D). No association with phenotypic trait(s) was reported in the original study for these two loci. But as the two lead SNPs are no longer genome-wide significant once conditioning on the SNPs established by independent studies after the publication of the original WTCCC study[18–20] we view them as 'missed' rather than 'novel' loci.

Furthermore, we identified two refined SNP associations: one is SNP rs11209026 in the exon of *IL23R* gene for association with Crohn's disease. It is predicted to be probably damaging by PolyPhen-2 and has a *P*-value of 1.41×10^{-21} compared with the previous best association *P*-value of 5.85×10^{-12} (SNP rs11805303 within intron 6 of *IL23R*, see Table 1) from the original WTCCC study. Although the lead SNP rs11209026 is still genome-wide significant after conditioning on the lead genotyped SNP, we consider it a refined signal for two reasons. First two SNPs reside within the same gene and in physical vicinity (30.4 kb apart). Second, the *P*-value of rs11209026 dropped by more than three orders of magnitude (from 4.2×10^{-21} to 7.6×10^{-17}), suggesting that the two signals are partially dependent or tagging the same underlying/untyped causal SNP(s) or haplotype(s). The other refined association is located in the *CUX2* region for association with T1D, where the newly identified top SNP rs1265564 has an association *P*-value of 1.68×10^{-16}. The original WTCCC study reported a best association *P*-value of 1.51×10^{-14} (rs17696736) within the gene *NAA25* (Table 1). These two SNPs are 780 kb apart, however, the *CUX2* gene has been shown to directly regulate the expression of NeuroD, a gene that can cause T1D when mutated.[21] The lead SNPs for these two refined loci (*IL23R* for CD and *CUX2* for T1D) are no longer genome-wide significant after conditioning on the nearest lead SNPs reported in literature to date.[18,22]

We tend to believe that the two missed loci, namely the *IL2RA* locus for T1D and *CDKN2B* locus for T2D, are the same as reported in later studies for two reasons. First, the two missed loci fall in the same region with the SNPs reported in post-2007 literature with physical distance ranging from 3.6 to 28.8 kb. Although the level of LD is largely moderate, given such close physical vicinity, it is hard to rule out the possibility that both our lead SNPs and the SNPs reported in literature are tagging the same (untested) SNP(s) or haplotype(s). Our second reason is that the *P*-values of our lead SNPs drop by an order of three or four, and all $>2.9\times10^{-5}$ once conditional on the SNPs reported in the post-2007 literature. For the *IL2RA* locus, more detailed haplotype and fine-mapping analyses would be required to fully appreciate the complex architecture of causal variants in this region.

The independent imputation and association analysis using BEAGLE and PLINK identified the exact same four SNPs showing best association signals in the four regions identified by MaCH and MACH2DAT, confirming our findings that two are novel and two are refined, except that the BEALGE/PLINK generated a *P*-value for the *CDKN2B* locus slightly above the 5×10^{-8} threshold. Both MaCH and BEAGLE have been recommended for practical use because of their user-friendly interface and computational efficiency.[23–26]

We adopted a conservative genome-wide significance threshold 2.5×10^{-8} to guard against false positives particularly given that we are testing seven phenotypic traits instead of a single one. The fact that we only have genome-wide significant signals from four well-established regions suggests that our conservative threshold fulfilled its purpose. Future studies may gain additional power with more sophisticated methods to control type-I error[27–30] or with methods that handle multiple related phenotypes.[31,32] On the other hand, for four out of the seven traits, our 1000 genomes-based imputation detected nothing novel on top of the original WTCCC study, suggesting that the potential power of imputation is limited by the genetic architecture of the trait(s) of interest and the genomic coverage of the GWAS genotyping panel used.

We present here an example where 1000 genomes-based imputation identifies both novel and refined signals. By using 1000 genomes-based imputation, we identified two SNPs that are genome-wide significantly associated with two of the seven traits in the WTCCC study, neither discovered in the original study with only genotyped SNPs. The two SNPs 'missed' from the original analysis serve as positive controls because the two residing regions were both established by other studies after the 2007 WTCCC publication. Our analysis also provided further insights into two regions identified in the original study by identifying SNPs that are either more significant or point to a biologically more plausible gene. Importantly, we had no other signals based on our conservative genome-wide significance threshold, suggesting that we have no inflated false discovery rates. Taken together, our findings suggest that applying 1000 genomes-based imputation to the large number of GWAS data sets existing nowadays has the potential both to identify novel disease-associated genetic variants and to advance our understanding in known regions by examining a much denser set of imputed variants. We believe that larger reference panels continuing to be released by the 1000 Genomes Project will benefit the community even more, by performing single-marker analysis as presented here or rare or structural variant analysis.[11,33]

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank Prof David P Strachan at the St George's University of London for commenting on an earlier version of this manuscript. We acknowledge the WTCC for making the data available. A portion of this research was conducted using the Linux Cluster for Genetic Analysis (LinGA-II) funded by the Robert Dawson Evans Endowment of the Department of Medicine at Boston University School of Medicine and Boston Medical Center. The effort of DE and AF was supported by the Deutsche Forschungsgemeinschaft (DFG), grant no. FR 2821/2-1, and the German Ministry of Education and Research (BMBF) through the National Genome Research Network (NGFN). This project received infrastructure support through the DFG cluster of excellence 'Inflammation at Interfaces'. YL is partially supported by the NIH grant R01-HG006292 and 3-R01-CA082659-11S1.

1 WTCCC: Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature* 2007; **447**: 661–678.
2 Orho-Melander M, Melander O, Guiducci C *et al*: Common missense variant in the glucokinase regulatory protein gene is associated with increased plasma triglyceride and C-reactive protein but lower fasting glucose concentrations. *Diabetes* 2008; **57**: 3112–3121.
3 de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF: Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* 2008; **17**: R122–R128.
4 Li Y, Willer C, Sanna S, Abecasis G: Genotype imputation. *Annu Rev Genomics Hum Genet* 2009; **10**: 387–406.
5 Marchini J, Howie B: Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010; **11**: 499–511.
6 Thorisson GA, Smith AV, Krishnan L, Stein LD: The International HapMap Project Web site. *Genome Res* 2005; **15**: 1592–1593.
7 Durbin RM, Abecasis GR, Altshuler DL *et al*: A map of human genome variation from population-scale sequencing. *Nature* 2010; **467**: 1061–1073.
8 Liu JZ, Tozzi F, Waterworth DM *et al*: Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet* 2010; **42**: 436–440.
9 Sanna S, Pitzalis M, Zoledziewska M *et al*: Variants within the immunoregulatory CBLB gene are associated with multiple sclerosis. *Nat Genet* 2010; **42**: 495–497.
10 Purcell S, Neale B, Todd-Brown K *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.
11 Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR: MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* 2010; **34**: 816–834.
12 Adzhubei IA, Schmidt S, Peshkin L *et al*: A method and server for predicting damaging missense mutations. *Nat Methods* 2010; **7**: 248–249.
13 Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA: Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 2004; **74**: 106–120.
14 Browning BL, Browning SR: A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* 2009; **84**: 210–223.
15 Devlin B, Roeder K: Genomic control for association studies. *Biometrics* 1999; **55**: 997–1004.
16 The International HapMap Consortium: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; **449**: 851–861.
17 Pe'er I, Yelensky R, Altshuler D, Daly MJ: Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol* 2008; **32**: 381–385.
18 Barrett JC, Clayton DG, Concannon P *et al*: Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet* 2009; **41**: 703–707.
19 Saxena R, Voight BF, Lyssenko V *et al*: Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 2007; **316**: 1331–1336.
20 Shea J, Agarwala V, Philippakis AA *et al*: Comparing strategies to fine-map the association of common SNPs at chromosome 9p21 with type 2 diabetes and myocardial infarction. *Nat Genet* 2011; **43**: 801–805.
21 Iulianella A, Sharma M, Durnin M, Vanden Heuvel GB, Trainor PA: Cux2 (Cutl2) integrates neural progenitor development with cell-cycle progression during spinal cord neurogenesis. *Development* 2008; **135**: 729–741.
22 Barrett JC, Hansoul S, Nicolae DL *et al*: Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* 2008; **40**: 955–962.
23 Nothnagel M, Ellinghaus D, Schreiber S, Krawczak M, Franke A: A comprehensive evaluation of SNP genotype imputation. *Hum Genet* 2009; **125**: 163–171.
24 Pei YF, Li J, Zhang L, Papasian CJ, Deng HW: Analyses and comparison of accuracy of different genotype imputation methods. *PLoS One* 2008; **3**: e3551.
25 Zheng J, Li Y, Abecasis GR, Scheet P: A comparison of approaches to account for uncertainty in analysis of imputed genotypes. *Genet Epidemiol* 2011; **35**: 102–110.
26 Pei YF, Zhang L, Li J, Deng HW: Analyses and comparison of imputation-based association methods. *PLoS One* 2010; **5**: e10827.

27 Nyholt DR: A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* 2004; **74**: 765–769.
28 Conneely KN, Boehnke M: So many correlated tests, so little time! Rapid adjustment of p values for multiple correlated tests. *Am J Hum Genet* 2007; **81**: 1158–1168.
29 Gao XY: Multiple testing corrections for imputed SNPs. *Genet. Epidemiol* 2011; **35**: 154–158.
30 Wen SH, Lu ZS: Factors affecting the effective number of tests in genetic association studies: a comparative study of three PCA-based methods. *J Hum Genet* 2011; **56**: 428–435.
31 Kullo IJ, de Andrade M, Boerwinkle E, McConnell JP, Kardia SL, Turner ST: Pleiotropic genetic effects contribute to the correlation between HDL cholesterol, triglycerides, and LDL particle size in hypertensive sibships. *Am J Hypertens* 2005; **18**: 99–103.
32 Avery CL, He Q, North KE *et al*: A phenomics-based strategy identifies loci on APOC1, BRAP, and PLCG1 associated with metabolic syndrome phenotype domains. *PLoS Genet* 2011; **7**: e1002322.
33 Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zollner S: Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. *Am J Hum Genet* 2010; **87**: 604–617.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)