nature genetics

A reference panel of 64,976 haplotypes for genotype imputation

Shane McCarthy[1,98], Sayantan Das[2,3,98], Warren Kretzschmar[4,98], Olivier Delaneau[5], Andrew R Wood[6], Alexander Teumer[7,8], Hyun Min Kang[2,3], Christian Fuchsberger[2,3], Petr Danecek[9], Kevin Sharp[10], Yang Luo[1], Carlo Sidore[11], Alan Kwong[2,3], Nicholas Timpson[12], Seppo Koskinen[13], Scott Vrieze[14,15], Laura J Scott[2,3], He Zhang[16], Anubha Mahajan[4], Jan Veldink[17], Ulrike Peters[18,19], Carlos Pato[20], Cornelia M van Duijn[21], Christopher E Gillies[22], Ilaria Gandin[23], Massimo Mezzavilla[24,25], Arthur Gilly[1], Massimiliano Cocca[23], Michela Traglia[26], Andrea Angius[11], Jeffrey C Barrett[1], Dorrett Boomsma[27], Kari Branham[28], Gerome Breen[29,30], Chad M Brummett[31], Fabio Busonero[11], Harry Campbell[32], Andrew Chan[33,34], Sai Chen[2,3,35,36], Emily Chew[37], Francis S Collins[38], Laura J Corbin[12], George Davey Smith[12], George Dedoussis[39], Marcus Dorr[40,41], Aliki-Eleni Farmaki[39], Luigi Ferrucci[42], Lukas Forer[43], Ross M Fraser[31], Stacey Gabriel[44], Shawn Levy[45], Leif Groop[46–48], Tabitha Harrison[18], Andrew Hattersley[49], Oddgeir L Holmen[50], Kristian Hveem[50], Matthias Kretzler[35,36,51], James C Lee[52,53], Matt McGue[54], Thomas Meitinger[55–57], David Melzer[58], Josine L Min[12], Karen L Mohlke[59], John B Vincent[60–62], Matthias Nauck[8,41], Deborah Nickerson[63], Aarno Palotie[44,64–68], Michele Pato[20], Nicola Pirastu[23], Melvin McInnis[69], J Brent Richards[70–72], Cinzia Sala[26], Veikko Salomaa[13], David Schlessinger[73], Sebastian Schoenherr[43], P Eline Slagboom[74], Kerrin Small[72], Timothy Spector[72], Dwight Stambolian[75], Marcus Tuke[6], Jaakko Tuomilehto[76–79], Leonard H Van den Berg[17], Wouter Van Rheenen[17], Uwe Volker[41,80], Cisca Wijmenga[81], Daniela Toniolo[26], Eleftheria Zeggini[1], Paolo Gasparini[23,25], Matthew G Sampson[22], James F Wilson[32,82], Timothy Frayling[6], Paul I W de Bakker[83,84], Morris A Swertz[81,85], Steven McCarroll[86,87], Charles Kooperberg[18], Annelot Dekker[17], David Altshuler[44,66,88–91], Cristen Willer[16,35,36], William Iacono[54], Samuli Ripatti[92], Nicole Soranzo[1,93,94], Klaudia Walter[1], Anand Swaroop[95], Francesco Cucca[11], Carl A Anderson[1], Richard M Myers[45], Michael Boehnke[2,3], Mark I McCarthy[4,96,97], Richard Durbin[1,99], Gonçalo Abecasis[2,3,99] & Jonathan Marchini[4,10,99] for the Haplotype Reference Consortium

We describe a reference panel of 64,976 human haplotypes at 39,235,157 SNPs constructed using whole-genome sequence data from 20 studies of predominantly European ancestry. Using this resource leads to accurate genotype imputation at minor allele frequencies as low as 0.1% and a large increase in the number of SNPs tested in association studies, and it can help to discover and refine causal loci. We describe remote server resources that allow researchers to carry out imputation and phasing consistently and efficiently.

Over the last decade, large-scale international collaborative efforts have created successively larger and more ancestrally diverse genetic variation resources. For example, in 2007, the International HapMap Project produced a haplotype reference panel of 420 haplotypes at 3.1 million SNPs in three continental populations[1]. More recently, the 1000 Genomes Project has produced a series of data sets built using low-coverage whole-genome sequencing, culminating in 2015 in a reference panel (1000GP3) of 5,008 haplotypes at over 88 million variants from 26 worldwide populations[2]. In addition, several other projects have collected low-coverage whole-genome sequencing data in large numbers of samples that could potentially also be used to build haplotype reference panels[3–5]. A major use of these resources has been to facilitate imputation of unobserved genotypes into genome-wide association study (GWAS) samples that have been assayed using relatively sparse genome-wide microarray chips. As reference panels have increased in number of haplotypes, SNPs and populations, genotype imputation accuracy has increased, allowing researchers to impute and test SNPs for association at ever lower minor allele frequencies (MAFs). A succession of methods developments has provided researchers with the tools to cope with these increasingly larger panels[6–11].

We formed the Haplotype Reference Consortium (HRC; see URLs) to bring together as many whole-genome sequencing data sets as possible to build a much larger combined haplotype reference panel.

A full list of affiliations appears at the end of the paper.

Received 21 December 2015; accepted 18 July 2016; published online 22 August 2016; doi:10.1038/ng.3643

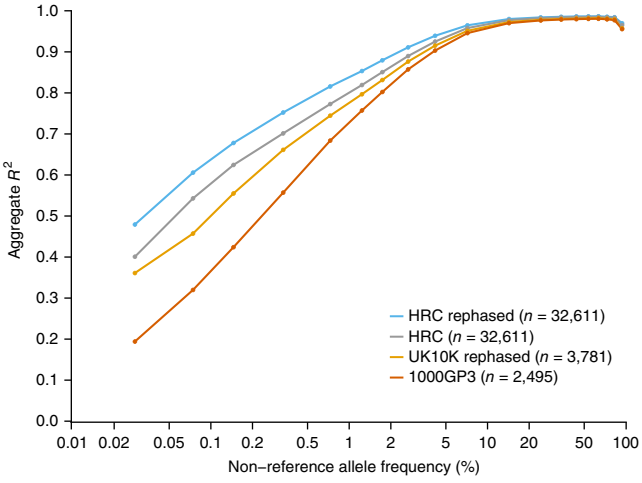

Figure 1 Performance of imputation using different reference panels. The *x* axis shows the non-reference allele frequency of the SNP being imputed on a log scale. The *y* axis shows imputation accuracy measured by aggregate R^2 value when imputing SNP genotypes into ten CEU samples. These results are based on using genotypes from sites on the Illumina Omni1M SNP array as pseudo-GWAS data.

By doing so, our aim is to provide a single centralized resource for human genetics researchers to carry out genotype imputation. Here we describe the first HRC reference panel that combines data sets from 20 different studies (**Supplementary Table 1**). The majority of these studies have low-coverage whole-genome sequencing data (4–8× coverage) and are known to consist of samples with predominantly European ancestry. However, the 1000 Genomes Project Phase 3 cohort, which has diverse ancestry, is also included. This reference panel consists of 64,976 haplotypes at 39,235,157 SNPs with evidence of having a minor allele count (MAC) greater or equal to 5.

We took the following approach to create the reference panel. We combined existing sets of genotype calls from each study to determine a 'union' set of 95,855,206 SNP sites with MAC ≥ 2. After initial tests, we decided for this first version of the HRC panel not to include small indels, as these were very inconsistently called across projects. We then used a standard tool to calculate consistently the genotype likelihoods for each sample at each site from the original study BAM files (Online Methods) and make a baseline set of genotype calls not based on linkage disequilibrium (LD). We next applied a number of filters to remove poor-quality sites (Online Methods). We restricted the site list to sites with MAC ≥ 5 on the basis of calls originally made by the individual studies, corresponding to a minimum MAF of 0.0077%, and then added back sites that are present on several commonly used SNP microarray chips for GWAS. Sites with lower MAF values would be likely to be poorly imputed. This site list consisting of 44,187,567 sites exhibited improved quality in comparison to the unfiltered site list with MAC ≥5 when assessed by measuring per-sample transition-to-transversion (Ts/Tv) ratio (**Supplementary Figs. 1** and **2**). We also detected and removed 301 duplicate samples across the whole data set (Online Methods).

Calling genotypes and phasing using low-coverage whole-genome sequencing data was a computationally challenging step for many of the 20 studies providing data. To reduce computation, we carried out this step on genotype likelihoods from all 32,611 samples together and leveraged the original separately called haplotypes from each study to help reduce the search space of the calling algorithm (Online Methods). We then applied a further refinement step in which the called genotypes were rephased using the SHAPEIT3 method[12], on the basis of experience from the UK10K project, which found that this rephasing approach produced substantially improved imputation accuracy when using the haplotypes[4]. After final genotype calling, we removed a further 123 samples (Online Methods) and filtered out 4,952,410 sites whose MAC values after refinement and sample removal were below 5, resulting in a final set of 39,235,157 sites and 32,488 samples. By measuring the genotype discordance of the called genotypes in comparison to Illumina Omni2.5M chip genotypes available for the 1000 Genomes Project samples, we showed that both our site filtering strategy and the increased sample size of the HRC panel led to improved accuracy (**Supplementary Table 2**). For example, we obtained non-reference allele discordance of 0.39% on the full HRC data set with site filtering, as compared to 0.67% on the subset of 1000GP3 samples.

We next carried out experiments to assess and illustrate downstream imputation performance in comparison to previous haplotype reference panels. To mimic a typical imputation analysis, we created a pseudo-GWAS data set using high-coverage Complete Genomics (CG) whole-genome sequencing genotypes for ten CEU (European-ancestry) samples (see URLs). We extracted the CG SNP genotypes at all the sites included on an Illumina 1M SNP array (Human1M-Duo v3C). These were used to impute the remaining genotypes, which were then compared to the held-out genotypes, stratifying results by MAF of the imputed sites. The HRC reference panel led to a large increase in imputation performance when using a 1M SNP chip, in comparison to 1000GP3 (R^2 = 0.64 versus 0.36 at MAF = 0.1%), and the rephasing step using SHAPEIT3 was also beneficial (**Fig. 1**). HRC imputation at 0.1% frequency provided similar accuracy to 1000GP3 imputation at 0.6% frequency. The results from a denser SNP chip (Illumina Omni 5M) and the sparser Illumina CoreExome are shown in **Supplementary Figures 3** and **4**.

To illustrate the benefits of using the HRC resource, we imputed a GWAS of 1,210 samples from the InCHIANTI study[13], including 534 samples that did not contribute to the HRC reference panel because they were not sequenced. Imputing using the HRC panel resulted in 15,501,516 SNPs passing an imputation quality threshold of $R^2 \geq 0.5$, in comparison to 13,238,968 variants (11,908,509 SNPs and 1,330,459 indels) when imputing using 1000 Genomes Project Phase 3 data, corresponding to an increase of over 2 million variants. Taking the intersection of the variant sites from the two panels to account for the filtering applied to the HRC panel resulted in 13,364,795 SNPs and 10,728,322 SNPs with $R^2 \geq 0.5$ for the HRC reference and 1000 Genomes Project Phase 3 panel, respectively. The majority of these additional SNPs occurred in the lower-frequency range (**Supplementary Table 3**).

We next tested the HRC-imputed genotypes for association with 93 circulating blood marker phenotypes, including many of relevance to human health such as lipids, vitamins, ions, inflammatory markers and adipokines[14,15]. This analysis highlighted potential new associations at the nominal GWAS significance threshold of $P < 5 \times 10^{-8}$ (**Supplementary Table 4**). When we repeated imputation using the HRC panel without the overlapping InCHIANTI samples, we obtained similar results (**Supplementary Table 4**). We took these SNPs forward for replication in SHIP and SHIP-TREND cohorts (Online Methods) and found that two of the SNPs replicated (**Supplementary Table 5**). Specifically, we found that SNP rs150956780 (MAF = 0.6%) was associated with the lactic dehydrogenase phenotype (meta-analysis *P* value = 3.779×10^{-29}) and SNP rs147142246 (MAF = 0.6%) was associated with the potassium phenotype (meta-analysis *P* value = 8.7×10^{-9}).

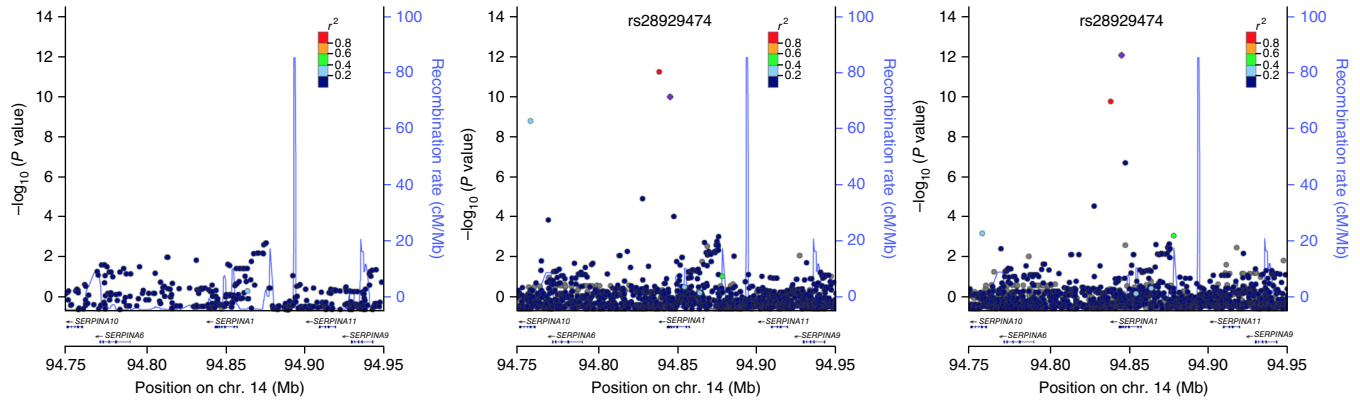

Figure 2 Association signal for the α_1-antitrypsin phenotype at the *SERPINA1* locus. Association test statistics on the $-\log_{10}$ (*P* value) scale (*y* axis) are plotted for each SNP position (*x* axis). Three different imputation panels were used: HapMap 2 (left), 1000GP3 (middle) and HRC release 1 (right). SNP rs28929474 is shown in purple, and other SNPs are colored according to the levels of LD (r^2) with this SNP.

We also found that it is possible for HRC-based imputation to refine signals of association. For example, association results using HapMap 2–, 1000GP3- and HRC-based imputation for the α_1-antitrypsin phenotype at the *SERPINA1* locus are shown in **Figure 2**. HRC-based imputation gave clear refinement of the signal at rare causal SNP rs28929474 (MAF = 0.5%) (**Supplementary Table 6**), known to predispose to the α_1-antitrypsin deficiency lung condition emphysema[16,17]. Similar results were obtained when using the HRC panel that excluded the InCHIANTI samples (data not shown).

As the HRC reference panel combines data from many different studies with a range of restrictions on data release, we have developed centralized imputation server resources (see URLs). Under this model, researchers upload phased or unphased genotype data and imputation is carried out on central servers. Once imputation is completed, researchers can download the imputed data sets. Along similar lines, we have also developed a lower-throughput phasing server for haplotype estimation of clinical samples that uses genotypes from high-coverage whole-genome sequencing data and takes advantage of rare variant sharing[18] (see URLs). It is our intention to make a limited subset of HRC haplotypes available for researchers via the European Genome-phenome Archive (EGA) for the sole purpose of phasing and imputation.

This first release of the HRC is the largest human genetic variation resource thus far and has been created via an unprecedented collaboration of data sharing across many groups. We envisage continuing to expand the HRC and are currently planning a second HRC release differing from the first release in two ways. First, we aim to substantially increase the ancestral diversity of the panel, by including data from sequencing studies in worldwide sample sets such as the CONVERGE study[19], AGVP[20] and HGDP[21]. Second, we aim to include indels in addition to SNP variants. At the limit of a reference panel consisting of the entire human population except the person being imputed, imputation would likely be almost perfect for alleles at any frequency, as the panel would contain close relatives who share long and almost identical tracts of sequence. Therefore, we do expect to be able to make future gains in imputation performance. In some populations that have experienced isolation (such as Sardinia or Iceland), we expect to approach this limit much faster. Thinking further ahead, we hope to work closely with efforts underway to collect high-coverage sequence for large numbers of samples such as the UK 100,000 Genomes Project (see URLs).

URLs. Haplotype Reference Consortium, http://www.haplotype-reference-consortium.org/; Michigan Imputation Server, https://imputationserver.sph.umich.edu/; Sanger Imputation Service, https://imputation.sanger.ac.uk/; Oxford Statistics Phasing Server, https://phasingserver.stats.ox.ac.uk/; genotype likelihood calculation scripts, https://github.com/mcshane/hrc-release1; GLPhase, https://github.com/wkretzsch/GLPhase; hapfuse, https://bitbucket.org/wkretzsch/hapfuse/src; Complete Genomics high-coverage whole-genome sequencing genotypes, http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130524_cgi_combined_calls/; 1000 Genomes Project Omni array genotypes, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/hd_genotype_chip/ALL.chip.omni_broad_sanger_combined.20140818.snps.genotypes.vcf.gz; 100,000 Genomes Project, http://www.genomicsengland.co.uk/the-100000-genomes-project/; GEMMA, http://www.xzlab.org/software.html; LocusZoom, http://locuszoom.sph.umich.edu/locuszoom/; 1000GP3 related samples, ftp://ftp.1000genomes.ebi.ac.uk//vol1/ftp/release/20130502/20140625_related_individuals.txt; SNP chip site lists, http://www.well.ox.ac.uk/~wrayner/strand/.

METHODS

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

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to all participants of all the studies that have contributed data to the HRC. J.M. acknowledges support from the ERC (grant 617306). W.K. acknowledges support from the Wellcome Trust (grant WT097307). S. McCarthy and R.D. acknowledge support from Wellcome Trust grant WT090851. A full list of acknowledgments for the cohorts is given in the **Supplementary Note**.

AUTHOR CONTRIBUTIONS

The HRC was initially conceived by discussions between J.M., G.A., R.D., M.I.M. and M.B. Analysis and methods development were carried out by S. McCarthy, S.D., W.K., O.D., A.R.W., P.D. and H.M.K. Supervision of the research was provided by J.M., G.A. and R.D. The Michigan Imputation Server was developed by C.F., L. Forer S.S. and G.A. The Sanger Imputation Service was developed by P.D., S. McCarthy and R.D. The Oxford Statistics Phasing Server was developed by W.K., K. Sharp and J.M. All other authors contributed data sets to the project or provided advice.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. International HapMap Consortium. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
2. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
3. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46**, 818–825 (2014).
4. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**, 8111 (2015).
5. Sidore, C. *et al.* Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat. Genet.* **47**, 1272–1281 (2015).
6. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
7. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
8. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).
9. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
10. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
11. Fuchsberger, C., Abecasis, G.R. & Hinds, D.A. minimac2: faster genotype imputation. *Bioinformatics* **31**, 782–784 (2015).
12. O'Connell, J. *et al.* Haplotype estimation for biobank-scale data sets. *Nat. Genet.* **48**, 817–820 (2016).
13. Ferrucci, L. *et al.* Subsystems contributing to the decline in ability to walk: bridging the gap between epidemiology and geriatric practice in the InCHIANTI study. *J. Am. Geriatr. Soc.* **48**, 1618–1625 (2000).
14. Melzer, D. *et al.* A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet.* **4**, e1000072 (2008).
15. Wood, A.R. *et al.* Imputation of variants from the 1000 Genomes Project modestly improves known associations and can identify low-frequency variant–phenotype associations undetected by HapMap based imputation. *PLoS One* **8**, e64343 (2013).
16. Bathurst, I.C., Travis, J., George, P.M. & Carrell, R.W. Structural and functional characterization of the abnormal Z α_1-antitrypsin isolated from human liver. *FEBS Lett.* **177**, 179–183 (1984).
17. Ferrarotti, I. *et al.* Serum levels and genotype distribution of α_1-antitrypsin in the general population. *Thorax* http://dx.doi.org/10.1136/thoraxjnl-2011-201321 (2012).
18. Sharp, K., Kretzschmar, W., Delaneau, O. & Marchini, J. Phasing for medical sequencing using rare variants and large haplotype reference panels. *Bioinformatics* **32**, 1974–1980 (2016).
19. CONVERGE Consortium. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* **523**, 588–591 (2015).
20. Gurdasani, D. *et al.* The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**, 327–332 (2015).
21. Rosenberg, N.A. *et al.* Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).

[1]Human Genetics, Wellcome Trust Sanger Institute, Hinxton, UK. [2]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA. [3]Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, USA. [4]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [5]Genetics and Development, University of Geneva, Geneva, Switzerland. [6]Genetics of Complex Traits, Institute of Biomedical Science, University of Exeter Medical School, Exeter, UK. [7]Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany. [8]DZHK (German Centre for Cardiovascular Research), Greifswald, Germany. [9]Vertebrate Resequencing Informatics, Wellcome Trust Sanger Institute, Hinxton, UK. [10]Department of Statistics, University of Oxford, Oxford, UK. [11]IRGB, CNR, Sardinia, Italy. [12]MRC Integrative Epidemiology Unit, University of Bristol, Oakfield Grove, UK. [13]THL, Helsinki, Finland. [14]Institute for Behavioral Genetics, University of Colorado, Boulder, Colorado, USA. [15]Department of Psychology and Neurosurgery, University of Colorado, Boulder, Colorado, USA. [16]Division of Cardiovascular Medicine, Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan, USA. [17]Department of Neurology and Neurosurgery, Brain Center Rudolf Magnus, Utrecht, the Netherlands. [18]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. [19]Department of Epidemiology, University of Washington School of Public Health, Seattle, Washington, USA. [20]Department of Psychiatry, SUNY Downstate, Brooklyn, New York, USA. [21]Genetic Epidemiology Unit, Department of Epidemiology, Erasmus MC, Rotterdam, the Netherlands. [22]Department of Pediatrics–Nephrology, University of Michigan School of Medicine, Ann Arbor, Michigan, USA. [23]Department of Medical, Surgical and Health Sciences, University of Trieste, Trieste, Italy. [24]Genetica Medica, IRCCS Burlo Garofolo, Trieste, Italy. [25]Department of Experimental Genetics, Sidra, Doha, Qatar. [26]Genetics and Cell Biology, San Raffaele Research Institute, Milan, Italy. [27]Netherlands Twin Register, Department of Biological Psychology, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands. [28]Department of Ophthalmology and Visual Sciences, University of Michigan, Ann Arbor, Michigan, USA. [29]MRC Social Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK. [30]NIHR Biomedical Research Centre for Mental Health, Institute of Psychiatry, Psychology and Neuroscience, King's College London and the South London Maudsley Hospital, London, UK. [31]Department of Anesthesiology, University of Michigan, Ann Arbor, Michigan, USA. [32]Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, UK. [33]Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA. [34]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA. [35]Department of Computational Medicine, University of Michigan, Ann Arbor, Michigan, USA. [36]Department of Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA. [37]Division of Epidemiology and Clinical Applications, National Eye Institute, Bethesda, Maryland, USA. [38]Medical Genomics and Metabolic Genetics Branch, National Human Genome Research Institute, US National Institutes of Health, Bethesda, Maryland, USA. [39]Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University, Athens, Greece. [40]Department of Internal Medicine B, University Medicine Greifswald, Greifswald, Germany. [41]Institute of Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Greifswald, Germany. [42]Longitudinal Studies Section, Clinical Research Branch, Gerontology Research Center, National Institute on Aging, Baltimore, Maryland, USA. [43]Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Medical University of Innsbruck, Innsbruck, Austria. [44]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [45]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama, USA. [46]Department of Clinical Sciences, Diabetes and Endocrinology, University of Lund, Malmö, Sweden. [47]Finnish Institute for Molecular Medicine, University of Helsinki, Helsinki, Finland. [48]Research Programs Unit, Diabetes and Obesity, University of Helsinki, Helsinki, Finland. [49]Institute of Biomedical and Clinical Research, University of Exeter Medical School, Exeter, UK. [50]Hunt Research Centre, Department of Public Health and General Practice, Norwegian University of Science and Technology, Levanger, Norway. [51]Department of Internal Medicine, University of Michigan School of Medicine, Ann Arbor, Michigan, USA. [52]Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK. [53]Department of Medicine, University of Cambridge School of Clinical Medicine, Addenbrooke's Hospital, Cambridge, UK. [54]Department of Psychology, University of Minnesota, Minneapolis, Minnesota, USA. [55]Institute of Human Genetics, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany. [56]Institute of Human Genetics, Technische Universität München, Munich, Germany. [57]DZHK (German Centre for Cardiovascular Research), Partner Site Munich Heart Alliance, Munich, Germany. [58]Epidemiology and Public Health, Institute of Biomedical and Clinical Science, University of Exeter Medical School, Exeter, UK. [59]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [60]Molecular Neuropsychiatry and Development Laboratory, Centre for Addiction and Mental Health, Toronto, Ontario, Canada. [61]Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada. [62]Institute of Medical Science, University of Toronto, Toronto, Ontario, Canada. [63]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [64]Institute for Molecular Medicine, FIMM, Helsinki, Finland. [65]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA. [66]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [67]Psychiatric and Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts, USA. [68]Department of Neurology, Massachusetts General Hospital, Boston, Massachusetts, USA. [69]Department of Psychiatry, University of Michigan, Ann Arbor, Michigan, USA. [70]Department of Medicine, McGill University, Montreal, Quebec, Canada. [71]Department of Human Genetics, McGill University, Montreal, Quebec, Canada. [72]Department of Twin Research and Genetic Epidemiology, King's College London, London, UK. [73]National Institute on Aging, US National Institutes of Health, Baltimore, Maryland, USA. [74]Molecular Epidemiology Section, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, the Netherlands. [75]Department of Ophthalmology, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [76]Chronic Disease Prevention Unit, National Institute for Health and Welfare, Helsinki, Finland. [77]Dasman Diabetes Institute,

Dasman, Kuwait. [78]Center for Vascular Prevention, Danube University Krems, Krems, Austria. [79]Diabetes Research Group, King Abdulaziz University, Jeddah, Saudi Arabia. [80]Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany. [81]Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. [82]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh, UK. [83]Medical Genetics, University Medical Center Utrecht, Utrecht, the Netherlands. [84]Department of Genetics, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, the Netherlands. [85]University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Groningen, the Netherlands. [86]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [87]Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts, USA. [88]Diabetes Research Center (Diabetes Unit), Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA. [89]Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA. [90]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [91]Vertex Pharmaceuticals, Boston, Massachusetts, USA. [92]Department of Public Health, University of Helsinki, Helsinki, Finland. [93]Department of Haematology, University of Cambridge, Cambridge, UK. [94]NIHR Blood and Transplant Unit (BTRU) in Donor Health and Genomics, University of Cambridge, Cambridge, UK. [95]Neurobiology–Neurodegeneration and Repair Laboratory, National Eye Institute, US National Institutes of Health, Bethesda, Maryland, USA. [96]Oxford Centre for Diabetes, Endocrinology and Metabolism, Radcliffe Department of Medicine, University of Oxford, Oxford, UK. [97]Oxford NIHR Biomedical Research Centre, Churchill Hospital, Headington, Oxford, UK. [98]These authors contributed equally to this work. [99]These authors jointly directed this work. Correspondence should be addressed to J.M. (marchini@stats.ox.ac.uk), G.A. (goncalo@umich.edu) or R.D. (rd@sanger.ac.uk).

ONLINE METHODS

Union site list. Every study provided us with the most recent version of their haplotypes in VCF format with one VCF for every autosome. For every cohort, bcftools (v0.2.0-rc12) was used to create a whole-autosome, SNP-only site list with alternate and total allele count information from these per-chromosome haplotypes. Multiallelic SNPs were divided into biallelic sites using 'bcftools norm'. The per-cohort site lists were merged into a single file that correctly merged alternate and total allele counts. We created site lists called MAC2 and MAC5 containing only sites with a MAC value across all studies of ≥2 and ≥5, respectively, using bcftools. These site lists contained 95,855,206 and 51,060,347 sites, respectively.

Genotype likelihood calculations. The 'samtools mpileup' command was used to generate genotype likelihoods at all MAC2 sites on a per-sample basis from each sample's BAM file. The pipeline and software versions have been made available online (see URLs). The resulting BCF files were merged using the 'bcftools merge' command, and MAC2 sites and alleles were extracted using the 'bcftools call' command. The use of 'bcftools call' here made a baseline set of non-LD-based genotype calls for each site across all samples. These calls were used for some initial sample quality control. We calculated genotype likelihoods on 33,070 samples in total.

Site filtering. We used an *ad hoc* method for initial variant filtering that enabled us to identify variants that had been filtered out 'quite often' by our submitting studies. For each site and for each cohort, we labeled the site as 'called' in that study if the putative calls from bcftools based on genotype likelihoods exhibited more than one allele in that cohort or 'not called' if the site showed no variation. We also used the haplotype sets provided by each study to determine whether each study had filtered out each site using their own internal calling pipeline. To determine a threshold of 'number of times filtered out', we stratified the sites according to their called status versus their filtered status (**Supplementary Fig. 5**). We also measured the Ts/Tv ratio of the set of SNPs for each of these stratified combinations. SNPs corresponding to the cells above the red line in **Supplementary Figure 5** were filtered out, removing all cells that had been filtered out by more than four studies or had a Ts/Tv ratio less than 1.7.

We also applied a set of additional site filters as follows. We filtered out sites not on the MAC5 site list to restrict the list to sites that could be well imputed. We also filtered out sites if in any study (apart from the 1000 Genomes Project) they had a Hardy–Weinberg equilibrium P value $<1 \times 10^{-10}$ or an overall inbreeding coefficient <-0.1. This coefficient (denoted f) was calculated as a departure from Hardy–Weinberg equilibrium by solving $Aa = 2pq - 2fpq$ (where p and q are the observed allele frequencies and Aa is the observed proportion of heterozygous calls). Negative values indicate an excess of heterozygous calls, and positive values indicate an excess of homozygosity. We also filtered out sites with MAF >0.1 that were called in fewer than three of the studies and were not called in the 1000 Genomes Project (the latter restriction kept sites present at high frequencies in non-European populations that were only called in the 1000 Genomes Project). We also filtered out sites called only in the GoNL study or IBD cohort. We completely excluded GPC haplotypes from this step of the site list creation process.

After applying these filters, the site list comprised 44,038,997 sites. Finally, we made sure that 4,914,335 sites found on a selection of commonly used SNP genotyping arrays and those used in the GIANT Consortium and the Global Lipids Consortium (**Supplementary Table 7**) were included in the final site list. The final site list after this filtering contained 44,187,567 sites.

Sample filtering. Having used 'bcftools call' to extract sites and alleles, we had a set of baseline non-LD-based genotype calls. On the basis of these calls for chromosome 22, some outlier samples were evident, and we removed 150 samples showing evidence of fewer than 10,000 non-reference SNPs or more than 10 singletons across the chromosome. This left a total of 32,920 samples.

To detect possible duplicates, we used the original genotype calls submitted by the individual studies. We selected 1,000 random sites that (i) were biallelic; (ii) had European MAF >5% in 1000GP3; and (iii) had no missing data in any of the individual studies. Using the 'bcftools gtcheck' command, we counted the number of genotypes that differed between each sample pair. There was a clear set of 269 sample pairs with very few genotypes differing over the 1,000 sites. We identified these samples as duplicates either within or between studies and removed one of the samples in each pair as described in **Supplementary Table 8**. Because some samples were represented more than twice, there were a total of 261 samples removed owing to duplicates. Before genotype calling, we also removed (i) 9 samples for which we had CG data, so that we could use these samples for testing purposes; (ii) 31 samples from 1000GP3 that were related (see URLs); and (iii) 8 samples from the HELIC, AMD and ProjectMinE studies with sample labeling inconsistencies. These filters resulted in 32,611 samples being used for the genotype calling and phasing steps.

In addition, after phasing, 83 samples from the AMD study were excluded as the consent for these samples had been removed. We also repeated the duplicate detection process on the final HRC genotype calls, as some studies increased in size late within the analysis process. This resulted in an additional 40 samples being removed and a total of 32,488 samples in the final phased reference panel.

Genotype calling method leveraging existing haplotype calls. We called genotypes from the genotype likelihoods computed on the HRC samples by extending the SNPTools[22] algorithm to leverage preexisting haplotypes available from each cohort. Like other phasing and calling approaches[8,10], SNPTools is a Markov chain Monte Carlo (MCMC) approach in which each sample's haplotypes and genotypes are iteratively updated using the current estimates of all other samples. A low-complexity hidden Markov model (HMM) with just four states is used to update each sample, where the states are a set of four 'surrogate parent' haplotypes. The MCMC sampler employs a Metropolis–Hastings step to sample the set of surrogate parents. In large sample sizes, the search space for these surrogate haplotypes is huge and results in low acceptance rates for the sampler. Our extension, called GLPhase (see URLs), uses preexisting haplotypes to restrict the set of possible haplotypes from which the Metropolis–Hastings sampler may choose surrogate parent haplotypes. For each individual, we restrict the search space to the 200 haplotypes that most closely match the 2 preexisting haplotypes for the individual using a Hamming distance metric (100 for each haplotype). We run the method on chunks of 1,024 sites at a time, which is the default setting for SNPTools. As the preexisting haplotypes from each study do not contain exactly the same set of sites, we filled in missing alleles in the preexisting haplotypes from our site list using the major allele at each site.

Restricting the search space in this way allows us to reduce the number of burn-in iterations from 56 to 5, the number of sampling iterations from 200 to 95, and the number of Metropolis–Hastings steps taken at each iteration for each individual from $2N$ to 100, where N is the number of samples being phased. This reduces the complexity of our phasing algorithm from $O(N^2)$ to $O(N)$. Although our implementation of the Hamming distance search has complexity $O(N^2)$, for $N = 30{,}000$, the impact of the search on run time is small (~5% of run time on each chunk). A chunk of 1,024 sites can be phased in ~200 min using ~1.3 GB of RAM. Once sample sizes are encountered where the Hamming distance search begins to dominate, our implementation could be replaced with $O(N \log N)$ clustering algorithms that we have implemented within the SHAPEIT3 algorithm[12].

To illustrate how important GLPhase was to genotype calling and phasing on such a large sample size, we carried out a comparison to Beagle 3.1, Beagle 4.1 and the original SNPTools method. We ran all four methods on five randomly selected 1,024-site chunks from chromosome 20 on the cluster using increasing sample sizes and measured run time. **Supplementary Figure 6** shows that GLPhase is approximately 100 times faster than the next fastest method at the full HRC sample size. We did not compare the accuracy of the methods because GLPhase is the only method it is feasible to run on data sets large enough to make meaningful comparisons. It may well be the case that gains in accuracy can be made over GLPhase, and we plan to investigate this for future releases of the HRC panel.

Final phasing and haplotype estimation. We estimated haplotypes from GLPhase genotype calls using SHAPEIT3 (ref. 12). Chromosomes were phased in chunks consisting of 16,000 variants plus 3,300 variants overlapping with neighboring chunks on either side. The non-default command line option -w 0.5 was

doi:10.1038/ng.3643

used for SHAPEIT3. Chunks were ligated using the hapfuse program (see URLs). SHAPEIT3 does not handle multiple variants at the same genomic coordinate, so multiallelic sites (SNPs with three or four alleles) were shifted by 1 or 2 bp for rephasing and then moved back to their original position after chunk ligation.

Evaluation of the genotype calling process. We tested the genotype calling process on data from chromosome 20 with different combinations of site lists and sample sets to assess both the effects of site filtering and the benefits of increasing sample size. We evaluated three different site lists: the 1000 Genomes Project Phase 3 set of sites (775,927), our HRC MAC5 site list (1,128,114) and our HRC MAC5 site list with additional site filtering (1,006,559). We ran the genotype calling method on three different sets of samples: the 2,525 original 1000 Genomes Project Phase 3 samples, a subset of 13,309 HRC samples that we used at an early stage of HRC testing (HRC Pilot) from the 1000GP3, AMD, GoNL, GoT2D, ORCADES, SardinIA, FINLAND and UK10K studies, and the near-final full set of 32,905 HRC samples. We called genotypes using GLPhase on each of these nine data sets and examined genotype discordance as compared to Illumina Omni2.5M genotypes produced by the 1000 Genomes Project. For this comparison, we focused only on genotypes from 365 samples shared across the three sample sets and at 42,244 SNP sites. We calculated percentage discordance for the three possible genotypes consisting of reference and alternate alleles as well as an overall non-reference allele discordance rate. Results are shown in **Supplementary Table 2**.

Downstream imputation performance. We assessed the imputation accuracy of four different reference panels: 1000 Genomes Project Phase 3, UK10K and two versions of the HRC reference panel, with and without rephasing with SHAPEIT3. To do this, we used high-coverage whole-genome sequencing data made publicly available by CG (see URLs). For the pseudo-GWAS samples, we used data from ten CEU samples that also occur in the 1000 Genomes Project Phase 3 samples. These samples were removed from the various reference panels before using them to assess imputation performance.

Three pseudo-GWAS panels were created on the basis of three chip lists (see URLs): the Illumina Omni5M SNP array (HumanOmni5-4v1-1_A), the Illumina Omni1M SNP array (Human1M-Duo v3C) and the Illumina CoreExome SNP array (humancoreexome-12v1-1_a). For these comparisons, we only used sites in the intersection of the reference panels to enable direct comparison.

These pseudo-chip genotypes were used to impute the remaining genotypes, which were then compared to the held-out genotypes, stratifying results by MAF of the imputed sites.

Imputation was carried out using IMPUTE2 (ref. 7), which chooses a custom reference panel for each study individual for each 2-Mb segment of the genome. We set the k_{hap} parameter of IMPUTE2 to 1,000. All other parameters were set to default values. We stratified imputed variants into allele frequency bins and calculated the squared correlation between the imputed allele dosages at variants in each bin and the masked CG genotypes (called aggregate R^2 values in **Fig. 1**). The non-reference allele frequency for each SNP was calculated from HRC release 1 genotype likelihoods at sites with MAC ≥5. **Figure 1** shows the results for the Illumina Omni1M chip. **Supplementary Figures 3** and **4** show the results from the Illumina CoreExome chip and the Illumina Omni5M chip, respectively.

Details of imputation, association testing and replication in the InCHIANTI study. A total of 1,210 individuals from the InCHIANTI study were genotyped using the Illumina Infinium HumanHap550 genotyping array[13,14]. Individuals were prephased using autosomal SNPs after filtering out SNPs with MAF <1%, Hardy–Weinberg equilibrium P value $<1 \times 10^{-4}$ and missingness >1%. SNPs were also removed if they could not be remapped to the GRCh37 (hg19) human reference genome. This filtering resulted in 483,991 SNPs available for prephasing. Phasing was performed locally using SHAPEIT2 (ref. 10).

Imputation was performed remotely using the Michigan Imputation Server (see URLs). A total of 39,235,157 SNPs and 47,045,346 variants were imputed from the HRC and 1000 Genomes Project Phase 3 (v5) reference panels, respectively. An imputation quality threshold of $R^2 > 0.5$ was subsequently applied to both imputation data sets before association testing. This resulted in 15,501,516 and 13,589,949 variants available for association analysis derived from HRC- and 1000 Genomes Project–based imputation, respectively.

Measures for a total of 93 circulating blood factors available in the InCHIANTI study were double inverse normalized, while adjusting for age and sex, before association testing[14,15]. Association analysis was performed using a linear mixed-model framework as implemented in GEMMA (see URLs). The association plots in **Figure 2** were produced using LocusZoom (see URLs).

We attempted to replicate the associations reported in **Supplementary Table 3** in the SHIP and SHIP-TREND cohorts[23]. The SHIP samples were genotyped using Affymetrix Genome-Wide Human SNP Array 6.0. The SHIP-TREND samples were genotyped using the Illumina Human Omni2.5 array. Before imputation, duplicate samples (identified using identity by state), samples with mismatch between reported and genotyped sex, and samples with a very high heterozygosity rate were excluded. Additionally, all monomorphic SNPs, SNPs with duplicate chromosomal positions, SNPs with Hardy–Weinberg equilibrium P values <0.0001 and SNPs with call rates <95% were filtered out. Imputation was performed on the Sanger Imputation Service (see URLs) against the HRC panel. In total, 4,070 SHIP samples and 986 SHIP-TREND samples were included in genotype imputation. Association analyses were conducted using SNPTEST v2.5.2 (ref. 24). A subset of the phenotypes with new associations was also analyzed within the Avon Longitudinal Study of Parents and Children (ALSPAC). This included measures of magnesium and potassium in cord blood and measures of free thyroxine (FT_4) and vitamin D both in children and pregnant women. These did not replicate (data not shown), although meta-analysis was not performed owing to high heterogeneity between samples.

22. Wang, Y., Lu, J., Yu, J., Gibbs, R.A. & Yu, F. An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. *Genome Res.* **23**, 833–842 (2013).
23. Völzke, H. *et al.* Cohort profile: the study of health in Pomerania. *Int. J. Epidemiol.* **40**, 294–307 (2011).
24. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).