ARTICLES

A haplotype map of the human genome

The International HapMap Consortium*

Inherited genetic variation has a critical but as yet largely uncharacterized role in human disease. Here we report a public database of common variation in the human genome: more than one million single nucleotide polymorphisms (SNPs) for which accurate and complete genotypes have been obtained in 269 DNA samples from four populations, including ten 500-kilobase regions in which essentially all information about common DNA variation has been extracted. These data document the generality of recombination hotspots, a block-like structure of linkage disequilibrium and low haplotype diversity, leading to substantial correlations of SNPs with many of their neighbours. We show how the HapMap resource can guide the design and analysis of genetic association studies, shed light on structural variation and recombination, and identify loci that may have been subject to natural selection during human evolution.

Despite the ever-accelerating pace of biomedical research, the root causes of common human diseases remain largely unknown, preventative measures are generally inadequate, and available treatments are seldom curative. Family history is one of the strongest risk factors for nearly all diseases—including cardiovascular disease, cancer, diabetes, autoimmunity, psychiatric illnesses and many others—providing the tantalizing but elusive clue that inherited genetic variation has an important role in the pathogenesis of disease. Identifying the causal genes and variants would represent an important step in the path towards improved prevention, diagnosis and treatment of disease.

More than a thousand genes for rare, highly heritable 'mendelian' disorders have been identified, in which variation in a single gene is both necessary and sufficient to cause disease. Common disorders, in contrast, have proven much more challenging to study, as they are thought to be due to the combined effect of many different susceptibility DNA variants interacting with environmental factors.

Studies of common diseases have fallen into two broad categories: family-based linkage studies across the entire genome, and population-based association studies of individual candidate genes. Although there have been notable successes, progress has been slow due to the inherent limitations of the methods; linkage analysis has low power except when a single locus explains a substantial fraction of disease, and association studies of one or a few candidate genes examine only a small fraction of the 'universe' of sequence variation in each patient.

A comprehensive search for genetic influences on disease would involve examining all genetic differences in a large number of affected individuals and controls. It may eventually become possible to accomplish this by complete genome resequencing. In the meantime, it is increasingly practical to systematically test common genetic variants for their role in disease; such variants explain much of the genetic diversity in our species, a consequence of the historically small size and shared ancestry of the human population.

Recent experience bears out the hypothesis that common variants have an important role in disease, with a partial list of validated examples including *HLA* (autoimmunity and infection)[1], *APOE4* (Alzheimer's disease, lipids)[2], Factor V^{Leiden} (deep vein thrombosis)[3], *PPARG* (encoding PPARγ; type 2 diabetes)[4,5], *KCNJ11* (type 2 diabetes)[6], *PTPN22* (rheumatoid arthritis and type 1 diabetes)[7,8], insulin (type 1 diabetes)[9], *CTLA4* (autoimmune thyroid disease, type 1 diabetes)[10], *NOD2* (inflammatory bowel disease)[11,12], complement factor H (age-related macular degeneration)[13–15] and *RET* (Hirschsprung disease)[16,17], among many others.

Systematic studies of common genetic variants are facilitated by the fact that individuals who carry a particular SNP allele at one site often predictably carry specific alleles at other nearby variant sites. This correlation is known as linkage disequilibrium (LD); a particular combination of alleles along a chromosome is termed a haplotype.

LD exists because of the shared ancestry of contemporary chromosomes. When a new causal variant arises through mutation—whether a single nucleotide change, insertion/deletion, or structural alteration—it is initially tethered to a unique chromosome on which it occurred, marked by a distinct combination of genetic variants. Recombination and mutation subsequently act to erode this association, but do so slowly (each occurring at an average rate of about 10^{-8} per base pair (bp) per generation) as compared to the number of generations (typically 10^4 to 10^5) since the mutational event.

The correlations between causal mutations and the haplotypes on which they arose have long served as a tool for human genetic research: first finding association to a haplotype, and then subsequently identifying the causal mutation(s) that it carries. This was pioneered in studies of the *HLA* region, extended to identify causal genes for mendelian diseases (for example, cystic fibrosis[18] and diastrophic dysplasia[19]), and most recently for complex disorders such as age-related macular degeneration[13–15].

Early information documented the existence of LD in the human genome[20,21]; however, these studies were limited (for technical reasons) to a small number of regions with incomplete data, and general patterns were challenging to discern. With the sequencing of the human genome and development of high-throughput genomic methods, it became clear that the human genome generally displays more LD[22] than under simple population genetic models[23], and that LD is more varied across regions, and more segmentally structured[24–30], than had previously been supposed. These observations indicated that LD-based methods would generally have great value (because nearby SNPs were typically correlated with many of their neighbours), and also that LD relationships would

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

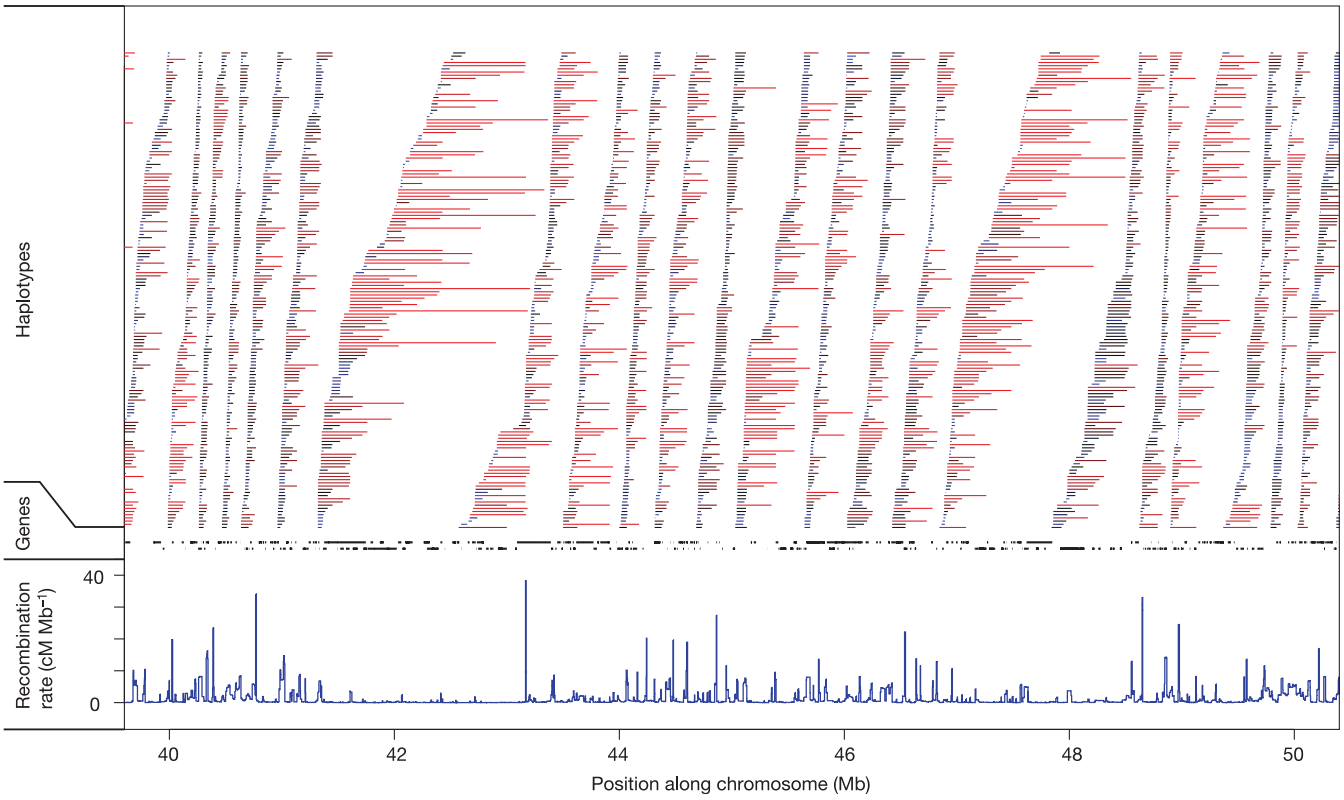


Figure 10 | The relationship among recombination rates, haplotype lengths and gene locations. Recombination rates in cM Mb^{-1} (blue). Non-redundant haplotypes with frequency of at least 5% in the combined sample (bars) and genes (black segments) are shown in an example gene-dense region of chromosome 19 (19q13). Haplotypes are coloured by the number of detectable recombination events they span, with red indicating many events and blue few.

ARTICLE

doi:10.1038/nature11632

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

By characterizing the geographic and functional spectrum of human genetic variation, the 1000 Genomes Project aims to build a resource to help to understand the genetic contribution to disease. Here we describe the genomes of 1,092 individuals from 14 populations, constructed using a combination of low-coverage whole-genome and exome sequencing. By developing methods to integrate information across several algorithms and diverse data sources, we provide a validated haplotype map of 38 million single nucleotide polymorphisms, 1.4 million short insertions and deletions, and more than 14,000 larger deletions. We show that individuals from different populations carry different profiles of rare and common variants, and that low-frequency variants show substantial geographic differentiation, which is further increased by the action of purifying selection. We show that evolutionary conservation and coding consequence are key determinants of the strength of purifying selection, that rare-variant load varies substantially across biological pathways, and that each individual contains hundreds of rare non-coding variants at conserved sites, such as motif-disrupting changes in transcription-factor-binding sites. This resource, which captures up to 98% of accessible single nucleotide polymorphisms at a frequency of 1% in related populations, enables analysis of common and low-frequency variants in individuals from diverse, including admixed, populations.

Recent efforts to map human genetic variation by sequencing exomes[1] and whole genomes[2–4] have characterized the vast majority of common single nucleotide polymorphisms (SNPs) and many structural variants across the genome. However, although more than 95% of common (>5% frequency) variants were discovered in the pilot phase of the 1000 Genomes Project, lower-frequency variants, particularly those outside the coding exome, remain poorly characterized. Low-frequency variants are enriched for potentially functional mutations, for example, protein-changing variants, under weak purifying selection[1,5,6]. Furthermore, because low-frequency variants tend to be recent in origin, they exhibit increased levels of population differentiation[6–8]. Characterizing such variants, for both point mutations and structural changes, across a range of populations is thus likely to identify many variants of functional importance and is crucial for interpreting individual genome sequences, to help separate shared variants from those private to families, for example.

We now report on the genomes of 1,092 individuals sampled from 14 populations drawn from Europe, East Asia, sub-Saharan Africa and the Americas (Supplementary Figs 1 and 2), analysed through a combination of low-coverage (2–6×) whole-genome sequence data, targeted deep (50–100×) exome sequence data and dense SNP genotype data (Table 1 and Supplementary Tables 1–3). This design was shown by the pilot phase[2] to be powerful and cost-effective in discovering and genotyping all but the rarest SNP and short insertion and deletion (indel) variants. Here, the approach was augmented with statistical methods for selecting higher quality variant calls from candidates obtained using multiple algorithms, and to integrate SNP, indel and larger structural variants within a single framework (see

Table 1 | Summary of 1000 Genomes Project phase I data

	Autosomes	Chromosome X	GENCODE regions*
Samples	1,092	1,092	1,092
Total raw bases (Gb)	19,049	804	327
Mean mapped depth (×)	5.1	3.9	80.3
SNPs			
No. sites overall	36.7 M	1.3 M	498 K
Novelty rate†	58%	77%	50%
No. synonymous/non-synonymous/nonsense	NA	4.7/6.5/0.097 K	199/293/6.3 K
Average no. SNPs per sample	3.60 M	105 K	24.0 K
Indels			
No. sites overall	1.38 M	59 K	1,867
Novelty rate†	62%	73%	54%
No. inframe/frameshift	NA	19/14	719/1,066
Average no. indels per sample	344 K	13 K	440
Genotyped large deletions			
No. sites overall	13.8 K	432	847
Novelty rate†	54%	54%	50%
Average no. variants per sample	717	26	39

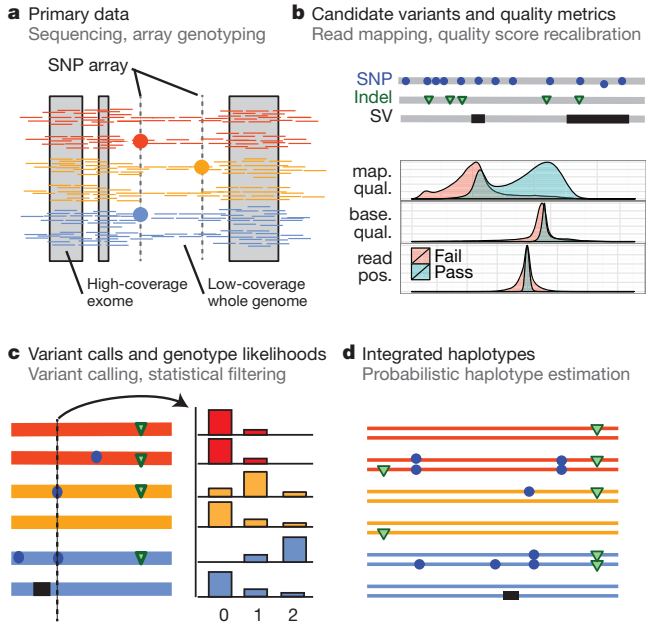
NA, not applicable.
* Autosomal genes only.
† Compared with dbSNP release 135 (Oct 2011), excluding contribution from phase I 1000 Genomes Project (or equivalent data for large deletions).

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

BOX 1

Constructing an integrated map of variation

The 1,092 haplotype-resolved genomes released as phase I by the 1000 Genomes Project are the result of integrating diverse data from multiple technologies generated by several centres between 2008 and 2010. The Box 1 Figure describes the process leading from primary data production to integrated haplotypes.



a, Unrelated individuals (see Supplementary Table 10 for exceptions) were sampled in groups of up to 100 from related populations (Wright's F_{ST} typically <1%) within broader geographical or ancestry-based groups[2]. Primary data generated for each sample consist of low-coverage (average 5×) whole-genome and high-coverage (average 80× across a consensus target of 24 Mb spanning more than 15,000 genes) exome sequence data, and high density SNP array information. **b**, Following read-alignment, multiple algorithms were used to identify candidate variants. For each variant, quality metrics were obtained, including information about the uniqueness of the surrounding sequence (for example, mapping quality (map. qual.)), the quality of evidence supporting the variant (for example, base quality (base. qual.) and the position of variant bases within reads (read pos.)), and the distribution of variant calls in the population (for example, inbreeding coefficient). Machine-learning approaches using this multidimensional information were trained on sets of high-quality known variants (for example, the high-density SNP array data), allowing variant sites to be ranked in confidence and subsequently thresholded to ensure low FDR. **c**, Genotype likelihoods were used to summarize the evidence for each genotype at bi-allelic sites (0, 1 or 2 copies of the variant) in each sample at every site. **d**, As the evidence for a single genotype is typically weak in the low-coverage data, and can be highly variable in the exome data, statistical methods were used to leverage information from patterns of linkage disequilibrium, allowing haplotypes (and genotypes) to be inferred.

Box 1 and Supplementary Fig. 1). Because of the challenges of identifying large and complex structural variants and shorter indels in regions of low complexity, we focused on conservative but high-quality subsets: biallelic indels and large deletions.

Overall, we discovered and genotyped 38 million SNPs, 1.4 million bi-allelic indels and 14,000 large deletions (Table 1). Several technologies were used to validate a frequency-matched set of sites to assess and control the false discovery rate (FDR) for all variant types. Where results were clear, 3 out of 185 exome sites (1.6%), 5 out of 281 low-coverage sites (1.8%) and 72 out of 3,415 large deletions (2.1%) could not be validated (Supplementary Information and Supplementary Tables 4–9). The initial indel call set was found to have a high FDR (27 out of 76), which led to the application of further filters, leaving an implied FDR of 5.4% (Supplementary Table 6 and Supplementary Information). Moreover, for 2.1% of low-coverage SNP and 18% of indel sites, we found inconsistent or ambiguous results, indicating that substantial challenges remain in characterizing variation in low-complexity genomic regions. We previously described the 'accessible genome': the fraction of the reference genome in which short-read data can lead to reliable variant discovery. Through longer read lengths, the fraction accessible has increased from 85% in the pilot phase to 94% (available as a genome annotation; see Supplementary Information), and 1.7 million low-quality SNPs from the pilot phase have been eliminated.

By comparison to external SNP and high-depth sequencing data, we estimate the power to detect SNPs present at a frequency of 1% in the study samples is 99.3% across the genome and 99.8% in the consensus exome target (Fig. 1a). Moreover, the power to detect SNPs at 0.1% frequency in the study is more than 90% in the exome and nearly 70% across the genome. The accuracy of individual genotype calls at heterozygous sites is more than 99% for common SNPs and 95% for SNPs at a frequency of 0.5% (Fig. 1b). By integrating linkage disequilibrium information, genotypes from low-coverage data are as accurate as those from high-depth exome data for SNPs with frequencies >1%. For very rare SNPs (≤0.1%, therefore present in one or two copies), there is no gain in genotype accuracy from incorporating linkage disequilibrium information and accuracy is lower. Variation among samples in genotype accuracy is primarily driven by sequencing depth (Supplementary Fig. 3) and technical issues such as sequencing platform and version (detectable by principal component analysis; Supplementary Fig. 4), rather than by population-level characteristics. The accuracy of inferred haplotypes at common SNPs was estimated by comparison to SNP data collected on mother–father–offspring trios for a subset of the samples. This indicates that a phasing (switch) error is made, on average, every 300–400 kilobases (kb) (Supplementary Fig. 5).

A key goal of the 1000 Genomes Project was to identify more than 95% of SNPs at 1% frequency in a broad set of populations. Our current resource includes ~50%, 98% and 99.7% of the SNPs with frequencies of ~0.1%, 1.0% and 5.0%, respectively, in ~2,500 UK-sampled genomes (the Wellcome Trust-funded UK10K project), thus

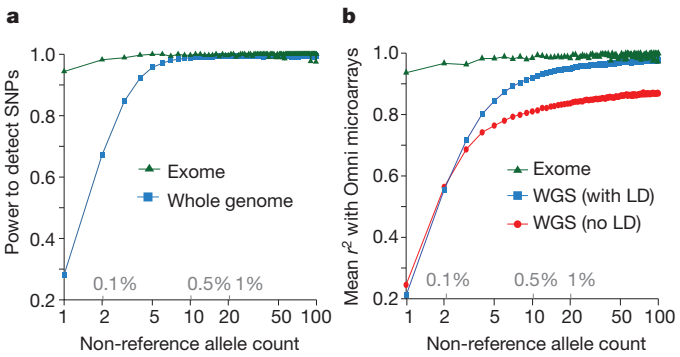


Figure 1 | Power and accuracy. **a**, Power to detect SNPs as a function of variant count (and proportion) across the entire set of samples, estimated by comparison to independent SNP array data in the exome (green) and whole genome (blue). **b**, Genotype accuracy compared with the same SNP array data as a function of variant frequency, summarized by the r^2 between true and inferred genotype (coded as 0, 1 and 2) within the exome (green), whole genome after haplotype integration (blue), and whole genome without haplotype integration (red). LD, linkage disequilibrium; WGS, whole-genome sequencing.

ARTICLE

OPEN

doi:10.1038/nature13997

The African Genome Variation Project shapes medical genetics in Africa

Deepti Gurdasani[1,2]*, Tommy Carstensen[1,2]*, Fasil Tekola-Ayele[3]*, Luca Pagani[1,4]*, Ioanna Tachmazidou[1]*, Konstantinos Hatzikotoulas[1], Savita Karthikeyan[1,2], Louise Iles[1,2,5], Martin O. Pollard[1], Ananyo Choudhury[6], Graham R. S. Ritchie[1,7], Yali Xue[1], Jennifer Asimit[1], Rebecca N. Nsubuga[8], Elizabeth H. Young[1,2], Cristina Pomilla[1,2], Katja Kivinen[1], Kirk Rockett[9], Anatoli Kamali[8], Ayo P. Doumatey[3], Gershim Asiki[8], Janet Seeley[8], Fatoumatta Sisay-Joof[10], Muminatou Jallow[10], Stephen Tollman[11,12], Ephrem Mekonnen[13], Rosemary Ekong[14], Tamiru Oljira[15], Neil Bradman[16], Kalifa Bojang[10], Michele Ramsay[6,17,18], Adebowale Adeyemo[3], Endashaw Bekele[19], Ayesha Motala[20], Shane A. Norris[21], Fraser Pirie[20], Pontiano Kaleebu[8], Dominic Kwiatkowski[1,9], Chris Tyler-Smith[1]§, Charles Rotimi[3]§, Eleftheria Zeggini[1]§ & Manjinder S. Sandhu[1,2]§

Given the importance of Africa to studies of human origins and disease susceptibility, detailed characterization of African genetic diversity is needed. The African Genome Variation Project provides a resource with which to design, implement and interpret genomic studies in sub-Saharan Africa and worldwide. The African Genome Variation Project represents dense genotypes from 1,481 individuals and whole-genome sequences from 320 individuals across sub-Saharan Africa. Using this resource, we find novel evidence of complex, regionally distinct hunter-gatherer and Eurasian admixture across sub-Saharan Africa. We identify new loci under selection, including loci related to malaria susceptibility and hypertension. We show that modern imputation panels (sets of reference genotypes from which unobserved or missing genotypes in study sets can be inferred) can identify association signals at highly differentiated loci across populations in sub-Saharan Africa. Using whole-genome sequencing, we demonstrate further improvements in imputation accuracy, strengthening the case for large-scale sequencing efforts of diverse African haplotypes. Finally, we present an efficient genotype array design capturing common genetic variation in Africa.

Globally, human populations show structured genetic diversity as a result of geographical dispersion, selection and drift. Understanding this variation can provide insights into evolutionary processes that shape both human adaptation and variation in disease susceptibility[1]. Although the Hapmap Project[2] and the 1000 Genomes Project[3] have greatly enhanced our understanding of genetic variation globally, the characterization of African populations remains limited. Other efforts examining African genetic diversity have been limited by variant density and sample sizes in individual populations[4], or have focused on isolated groups, such as hunter gatherers (HG)[5,6], limiting relevance to more widespread populations across Africa.

The African Genome Variation Project (AGVP) is an international collaboration that expands on these efforts by systematically assessing genetic diversity among 1,481 individuals from 18 ethno-linguistic groups from sub-Saharan Africa (SSA) (Fig. 1 and Supplementary Methods Tables 1 and 2) with the HumanOmni2.5M genotyping array and whole-genome sequences (WGS) from 320 individuals (Supplementary Methods Table 2). Importantly, the AGVP has evolved to help develop local resources for public health and genomic research, including strengthening research capacity, training, and collaboration across the region. We envisage that data from this project will provide a global resource for researchers, as well as facilitate genetic studies in Africa[7].

Population structure in SSA

On examining ~2.2 million variants, we found modest differentiation among SSA populations (mean pairwise F_{ST} 0.019) (Supplementary Methods and Supplementary Table 1). Differentiation among the Niger-Congo language groups—the predominant linguistic grouping across Africa was noted to be modest (mean pairwise F_{ST} 0.009) (Supplementary Table 1), providing evidence for the 'Bantu expansion'—a recent population expansion and movement throughout SSA originating in West Africa around 3,000 to 5,000 years ago[8].

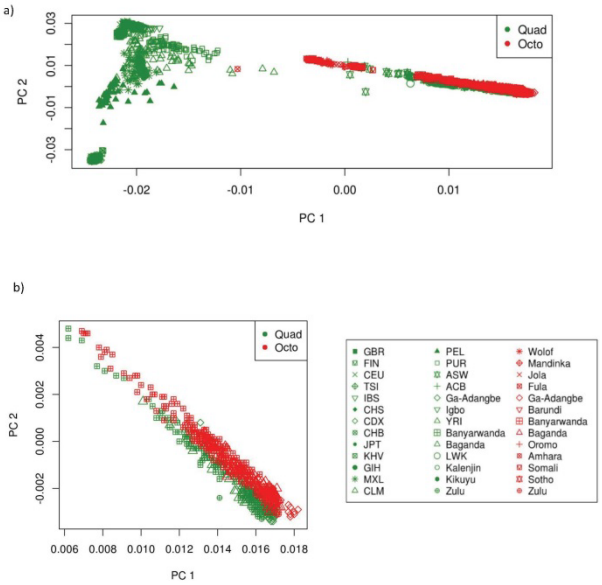
We identified 29.8 million single-nucleotide polymorphisms (SNPs) from Ethiopian, Zulu and Bagandan WGS (Extended Data Fig. 1 and

[1]Wellcome Trust Sanger Institute, Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [2]Department of Public Health and Primary Care, University of Cambridge, 2 Wort's Causeway, Cambridge, CB1 8RN, UK. [3]Centre for Research on Genomics and Global Health, National Human Genome Research Institute, National Institutes of Health, 12 South Drive, MSC 5635, Bethesda, Maryland 20891-5635, USA. [4]Department of Biological, Geological and Environmental Sciences, University of Bologna, Via Selmi 3, 40126 Bologna, Italy. [5]Department of Archaeology, University of York, King's Manor, York YO1 7EP, UK. [6]Sydney Brenner Institute of Molecular Bioscience (SBIMB), University of the Witwatersrand, The Mount, 9 Jubilee Road, Parktown 2193, Johannesburg, Gauteng, South Africa. [7]Vertebrate Genomics, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [8]Medical Research Council/Uganda Virus Research Institute, Plot 51-57 Nakiwogo Road, Uganda. [9]Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Headington, Oxford OX3 7BN, UK. [10]Medical Research Council Unit, Atlantic Boulevard, Serrekunda, PO Box 273, Banjul, The Gambia. [11]Medical Research Council/Wits Rural Public Health and Health Transitions Unit, School of Public Health, Education Campus, 27 St Andrew's Road, Parktown 2192, Johannesburg, Gauteng, South Africa. [12]INDEPTH Network, 38/40 Mensah Wood Street, East Legon, PO Box KD 213, Kanda, Accra, Ghana. [13]Institute of Biotechnology, Addis Ababa University, Entoto Avenue, Arat Kilo, 16087 Addis Ababa, Ethiopia. [14]Department of Genetics Evolution and Environment, University College, London, Gower Street, London WC1E 6BT, UK. [15]University of Haramaya, Department of Biology, PO Box 138, Dire Dawa, Ethiopia. [16]Henry Stewart Group, 28/30 Little Russell Street, London WC1A 2HN, UK. [17]Division of Human Genetics, National Health Laboratory Service, C/O Hospital and de Korte Streets, Braamfontein 2000, Johannesburg, South Africa. [18]School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Braamfontein 2000, Johannesburg, South Africa. [19]Department of Microbial, Cellular and Molecular Biology, College of Natural Sciences, Arat Kilo Campus, Addis Ababa University, PO Box 1176, Addis Ababa, Ethiopia. [20]Department of Diabetes and Endocrinology, University of KwaZulu-Natal, 719 Umbilo Road, Congella, Durban 4013, South Africa. [21]Department of Paediatrics, University of Witwatersrand, 7 York Road, Parktown 2198, Johannesburg, Gauteng, South Africa.

*These authors contributed equally to this work.

§These authors jointly supervised this work.

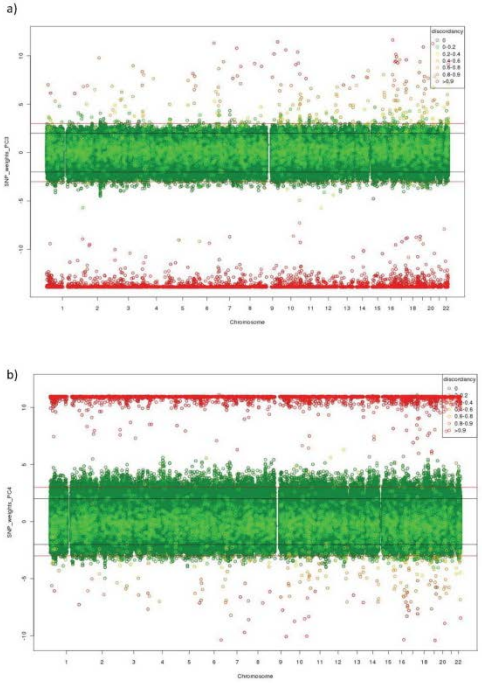
SN1. Figure 2: Chip effects apparent on global dataset principal component analysis- a representation of the top 8 PCs



SN1 Figure 2a shows global samples represented along PCs 1 and 2. Samples genotyped on the quad chip are shown in green, while those genotyped on the octo platform are shown in red. SN1 Figure 2b depicts PC 1 and 2 for samples only from the four populations that were genotyped across both chips. While slight separation is seen along PC2, this does not seem to primarily represent chip effects.

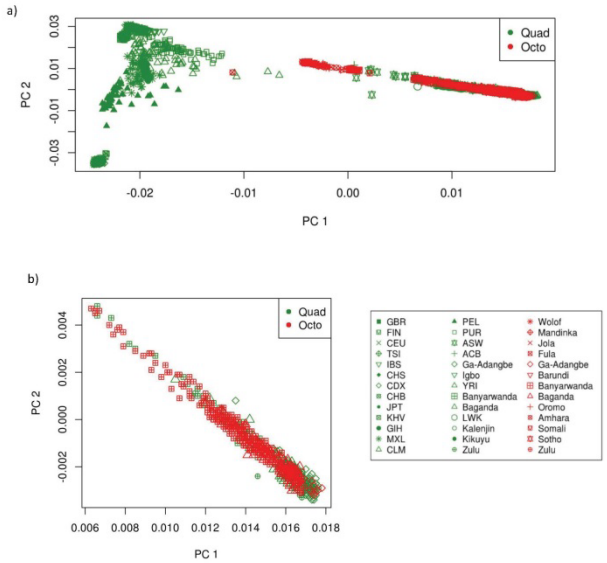
Supplementary Data 1: before, how, after

SN1. Figure 3: SNP weights for principal component 3 and 4 for the global dataset annotated with discordant SNPs in Baganda



SN1 Figure 3a and 3b represent standardised SNP loadings along PCs 3 and 4 for the global dataset along chromosomes 1-22. The black and red lines represent 2 and 3 SD thresholds from the mean respectively. Sites along chromosomes are coloured by the level of discordancy in genotypes between quad and octo platforms for 26 Baganda sample duplicates genotyped on both chips. There is a strong correlation observed between SNP weights and discordancy in genotypes among the two chips (Pearson's correlation r=0.77 and 0.61 for PCs 3 and 4 respectively).

SN1 Figure 4: PCA plots of global dataset after removal of SNPs with weight >3 SD from mean along PCs 3 and 4



SN1. Figure 4a shows global samples represented along PCs 1 and 2 after removal of SNPs with weights > 3 SD from the mean along PCs 3 and 4. Samples genotyped on the quad chip are shown in green, while those genotyped on the octo platform are shown in red. SN1 Figure 4b depicts PCs 1 and 2 for samples only from the four populations that were genotyped across both chips. No separation is observed by chip.

OPEN ACCESS Freely available online

PLOS | GENETICS

Imputation-Based Meta-Analysis of Severe Malaria in Three African Populations

Gavin Band[1], Quang Si Le[1], Luke Jostins[2], Matti Pirinen[1], Katja Kivinen[2], Muminatou Jallow[3,4], Fatoumatta Sisay-Joof[3], Kalifa Bojang[3], Margaret Pinder[3], Giorgio Sirugo[3], David J. Conway[3,5], Vysaul Nyirongo[6], David Kachala[6], Malcolm Molyneux[6,7], Terrie Taylor[8], Carolyne Ndila[9], Norbert Peshu[9], Kevin Marsh[9], Thomas N. Williams[9], Daniel Alcock[2], Robert Andrews[2], Sarah Edkins[2], Emma Gray[2], Christina Hubbart[1], Anna Jeffreys[1], Kate Rowlands[1], Kathrin Schuldt[1,10], Taane G. Clark[1,2,5], Kerrin S. Small[1,11], Yik Ying Teo[1,12], Dominic P. Kwiatkowski[1,2], Kirk A. Rockett[1,2], Jeffrey C. Barrett[2], Chris C. A. Spencer[1]*, Malaria Genomic Epidemiological Network¶

1 Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom, **2** Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom, **3** Medical Research Council Unit, Fajara, The Gambia, **4** Royal Victoria Teaching Hospital, Banjul, The Gambia, **5** London School of Hygiene and Tropical Medicine, London, United Kingdom, **6** Malawi-Liverpool-Wellcome Trust Clinical Research Programme, College of Medicine, University of Malawi, Blantyre, Malawi, **7** Liverpool School of Tropical Medicine, Liverpool, United Kingdom, **8** Blantyre Malaria Project, College of Medicine, University of Malawi, Blantyre, Malawi, **9** KEMRI–Wellcome Trust Research Programme, Kilifi, Kenya, **10** Department of Molecular Medicine, Bernhard Nocht Institute for Tropical Medicine, Hamburg, Germany, **11** Department of Twin Research and Genetic Epidemiology, King's College London, London, United Kingdom, **12** Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore

Abstract

Combining data from genome-wide association studies (GWAS) conducted at different locations, using genotype imputation and fixed-effects meta-analysis, has been a powerful approach for dissecting complex disease genetics in populations of European ancestry. Here we investigate the feasibility of applying the same approach in Africa, where genetic diversity, both within and between populations, is far more extensive. We analyse genome-wide data from approximately 5,000 individuals with severe malaria and 7,000 population controls from three different locations in Africa. Our results show that the standard approach is well powered to detect known malaria susceptibility loci when sample sizes are large, and that modern methods for association analysis can control the potential confounding effects of population structure. We show that pattern of association around the haemoglobin S allele differs substantially across populations due to differences in haplotype structure. Motivated by these observations we consider new approaches to association analysis that might prove valuable for multicentre GWAS in Africa: we relax the assumptions of SNP–based fixed effect analysis; we apply Bayesian approaches to allow for heterogeneity in the effect of an allele on risk across studies; and we introduce a region-based test to allow for heterogeneity in the location of causal alleles.

Citation: Band G, Le QS, Jostins L, Pirinen M, Kivinen K, et al. (2013) Imputation-Based Meta-Analysis of Severe Malaria in Three African Populations. PLoS Genet 9(5): e1003509. doi:10.1371/journal.pgen.1003509

Editor: Paul I. W. de Bakker, Brigham and Women's Hospital, United States of America

Received August 24, 2012; **Accepted** March 28, 2013; **Published** May 23, 2013

Funding: The MalariaGEN Project was supported by the Wellcome Trust (http://www.wellcome.ac.uk/) (WT077383/Z/05/Z) and the Bill and Melinda Gates Foundation (http://www.grandchallenges.org/) through the Foundations of the National Institutes of Health (http://www.nih.gov/) (566) as part of the Grand Challenges in Global Health Initiative. The Resource Centre for Genomic Epidemiology of Malaria is supported by the Wellcome Trust (090770/Z/09/Z). This research was supported by the Medical Research Council (G0600718, G0600230) and the Wellcome Trust Biomedical Ethics Enhancement Award (087285) and Strategic Award (096527). DP Kwiatkowski receives support from the Medical Research Council (http://www.mrc.ac.uk/index.htm) (G19/9). The Wellcome Trust also provide core awards to Wellcome Trust Centre for Human Genetics (075491/Z/04; 090532/Z/09/Z) and the Wellcome Trust Sanger Institute (077012/Z/05/Z). CCA Spencer was supported by a Wellcome Trust Career Development Fellowship [097364/Z/11/Z]. JC Barrett is supported under Wellcome Trust grant number WT098051. TN Williams was supported under a Wellcome Trust Senior Fellowship (091758/Z/10/Z). This paper was published with the permission of the Director of KEMRI. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

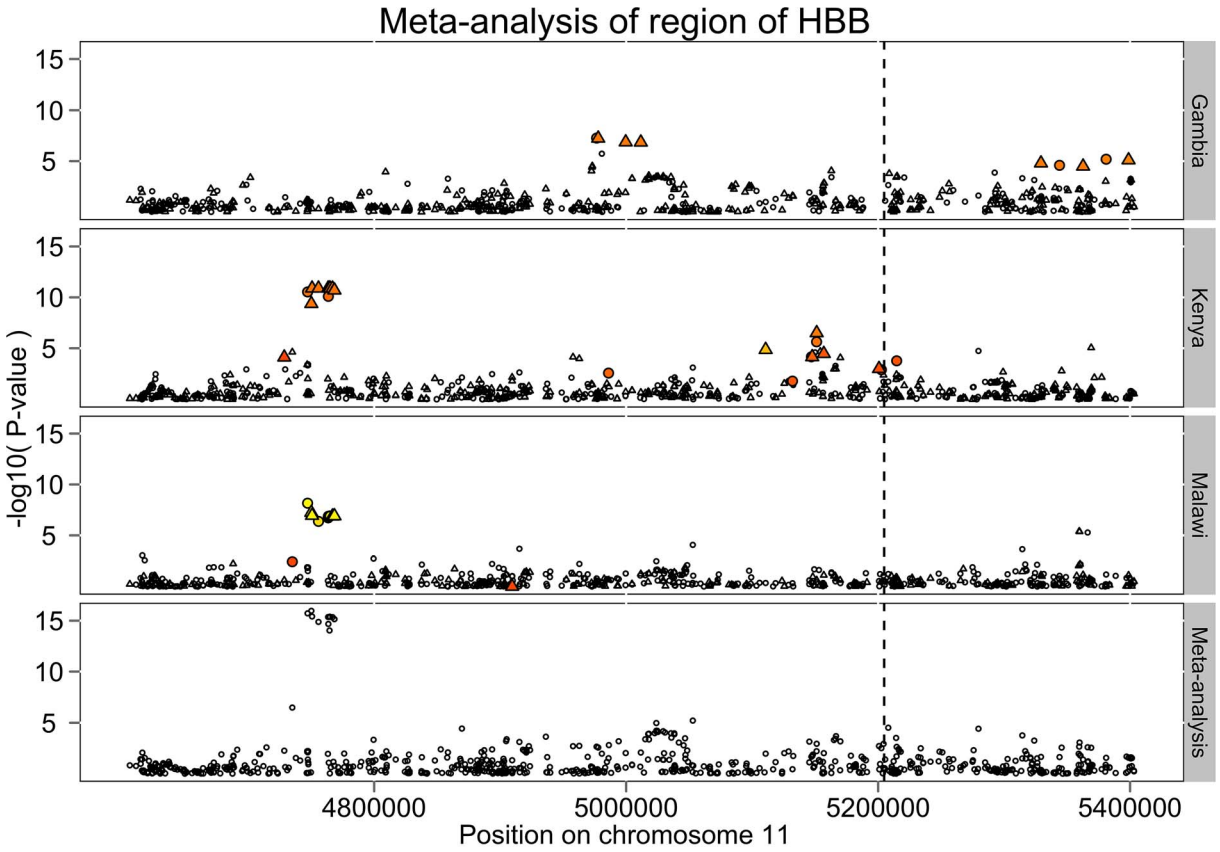
* E-mail: chris.spencer@well.ox.ac.uk

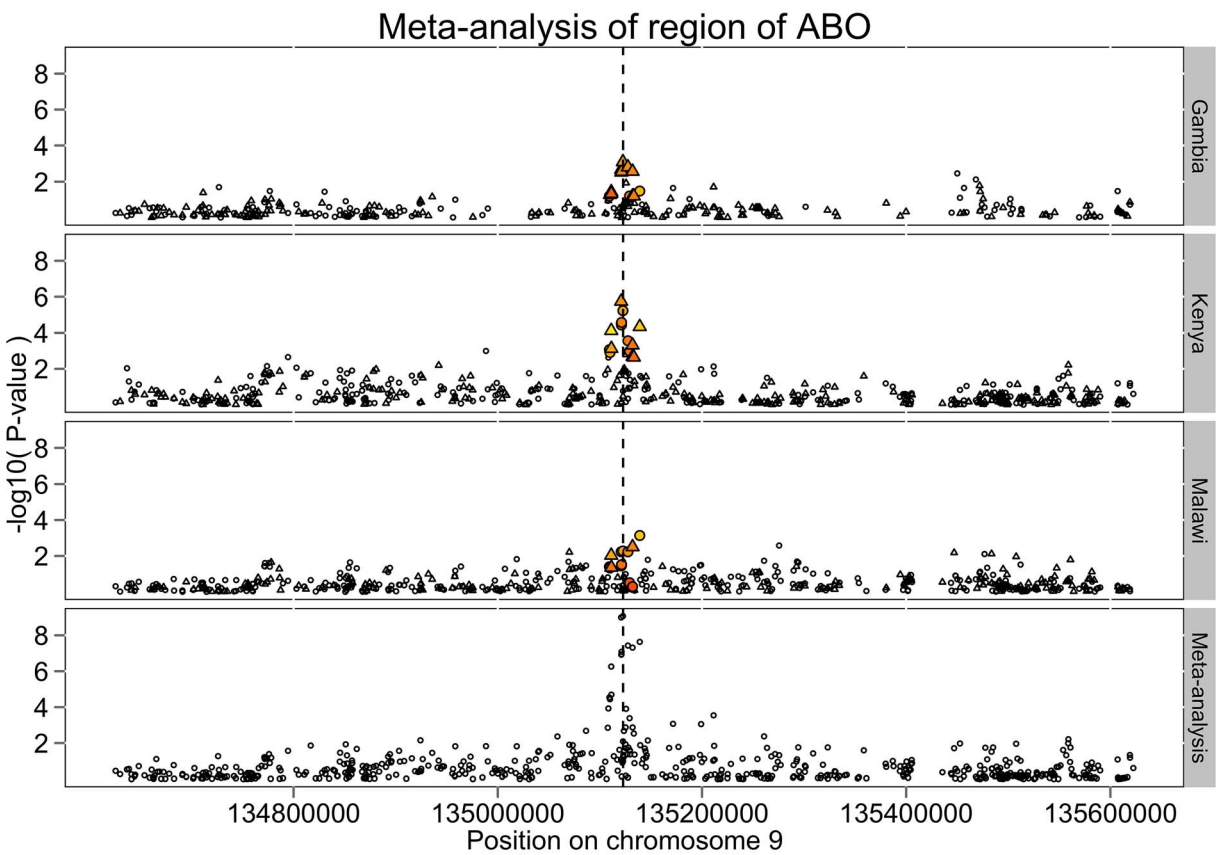
¶ For information about the Malaria Genomic Epidemiology Network (MalariaGEN) see www.malariagen.net.

Introduction

Severe malaria, meaning life-threatening complications of *Plasmodium falciparum* infection, kills on the order of a million African children each year [1]. However this represents only a small proportion of the total number of infected individuals, the majority of whom recover without life-threatening complications. Understanding the genetic basis of resistance to severe malaria could provide valuable insights into molecular mechanisms of pathogenesis and protective immunity that will aid the development of treatments and vaccines. It might also identify selective pressures that have shaped human physiology and susceptibility to other common diseases, because of the historical impact of malaria as a major cause of mortality in ancestral human populations.

Genome-wide association studies (GWAS) have identified thousands of genetic variants which predispose individuals to particular disease phenotypes. However, the vast majority of these studies are of non-communicable disease in collections of


Meta-analysis of region of HBB
-log10(P-value)
Gambia
Kenya
Malawi
Meta-analysis
15
10
5
4800000
5000000
5200000
5400000
Position on chromosome 11


Meta-analysis of region of ABO
-log10(P-value)
Gambia
Kenya
Malawi
Meta-analysis
8
6
4
2
134800000
135000000
135200000
135400000
135600000
Position on chromosome 9

nature genetics

Reappraisal of known malaria resistance loci in a large multicenter study

Malaria Genomic Epidemiology Network*

Many human genetic associations with resistance to malaria have been reported, but few have been reliably replicated. We collected data on 11,890 cases of severe malaria due to *Plasmodium falciparum* and 17,441 controls from 12 locations in Africa, Asia and Oceania. We tested 55 SNPs in 27 loci previously reported to associate with severe malaria. There was evidence of association at $P < 1 \times 10^{-4}$ with the *HBB*, *ABO*, *ATP2B4*, *G6PD* and *CD40LG* loci, but previously reported associations at 22 other loci did not replicate in the multicenter analysis. The large sample size made it possible to identify authentic genetic effects that are heterogeneous across populations or phenotypes, with a striking example being the main African form of G6PD deficiency, which reduced the risk of cerebral malaria but increased the risk of severe malarial anemia. The finding that G6PD deficiency has opposing effects on different fatal complications of *P. falciparum* infection indicates that the evolutionary origins of this common human genetic disorder are more complex than previously supposed.

It was recognized over half a century ago that malaria has been a major force of evolutionary selection on the human genome and that certain hematological disorders have risen to high frequency in malaria-endemic areas because they reduce the risk of death due to malaria[1–3]. Sickle hemoglobin (HbS) and glucose-6-phosphate dehydrogenase (G6PD) deficiency are often-quoted examples of natural selection due to malaria, and many other genetic associations with resistance or susceptibility to malaria have been reported[2–9]. However, the current literature contains many conflicting lines of evidence based on relatively small studies whose results have not been independently replicated.

To address this problem, we conducted a large multicenter case-control study of severe malaria across 12 locations in Burkina Faso, Cameroon, The Gambia, Ghana, Kenya, Malawi, Mali, Nigeria, Tanzania, Vietnam and Papua New Guinea (**Supplementary Fig. 1** and **Supplementary Table 1**). The structure of this consortial project has been described elsewhere[10], and information about each of the partner studies can be found on the Malaria Genomic Epidemiology Network (MalariaGEN) website (see URLs). We used the World Health Organization (WHO) definition of severe malaria, which comprises a broad spectrum of life-threatening clinical complications of *P. falciparum* infection[11–15]. In this report, we examine genetic associations with severe malaria in general and with two distinct clinical forms of severe malaria: cerebral malaria with a Blantyre coma score of less than 3 and severe malarial anemia with a hemoglobin level of less than 5 g/dl or a hematocrit level of less than 15%.

RESULTS

Samples and clinical data

The first stage of work was to collect standardized clinical data on severe malaria from multiple locations (**Supplementary Table 2**). This effort presented many practical challenges, as severe malaria is an acute illness that mainly occurs in resource-poor settings where laboratory facilities are limited and medical records can be unreliable. It was necessary to allow for variations in the design and implementation of the study in different settings, with study characteristics depending on a range of local circumstances. Investigators at different sites agreed at the outset on principles for sharing data and on standardized clinical definitions, and they also worked together to define best ethical practices across different local settings, including the development of guidelines for informed consent[10,16,17]. A set of web tools was developed to enable investigators to curate data in their locally used format before transforming them to the standardized format necessary for data from different sites to be merged.

After data curation and quality control (Online Methods), 11,890 cases of severe malaria and 17,441 controls were included for analysis (**Table 1** and **Supplementary Table 3**). Controls were intended to be representative of the populations to which the cases belonged; that is, a minority of controls may have subsequently gone on to develop severe malaria. The ancestry composition of the cases and controls at each location is shown in **Supplementary Table 4**. A total of 6,283 cases had cerebral malaria or severe malarial anemia, of which 3,345 had cerebral malaria only, 2,196 had severe malarial anemia only and 742 had both cerebral malaria and severe malarial anemia (**Table 1**). A further 5,607 cases did not have cerebral malaria or severe malarial anemia according to the criteria used here but satisfied the WHO definition of severe malaria, which includes a range of other clinical complications such as acidosis, respiratory distress and hypoglycemia that are not explored in detail in the present analysis[11].

Most of the cases of severe malaria were young children, with median ages ranging from 1.3 to 3.8 years at different study sites, except in Vietnam where most cases were young adults with a median

*A full list of authors and affiliations appears at the end of the paper.

Received 25 November 2013; accepted 4 September 2014; published online 28 September 2014; doi:10.1038/ng.3107

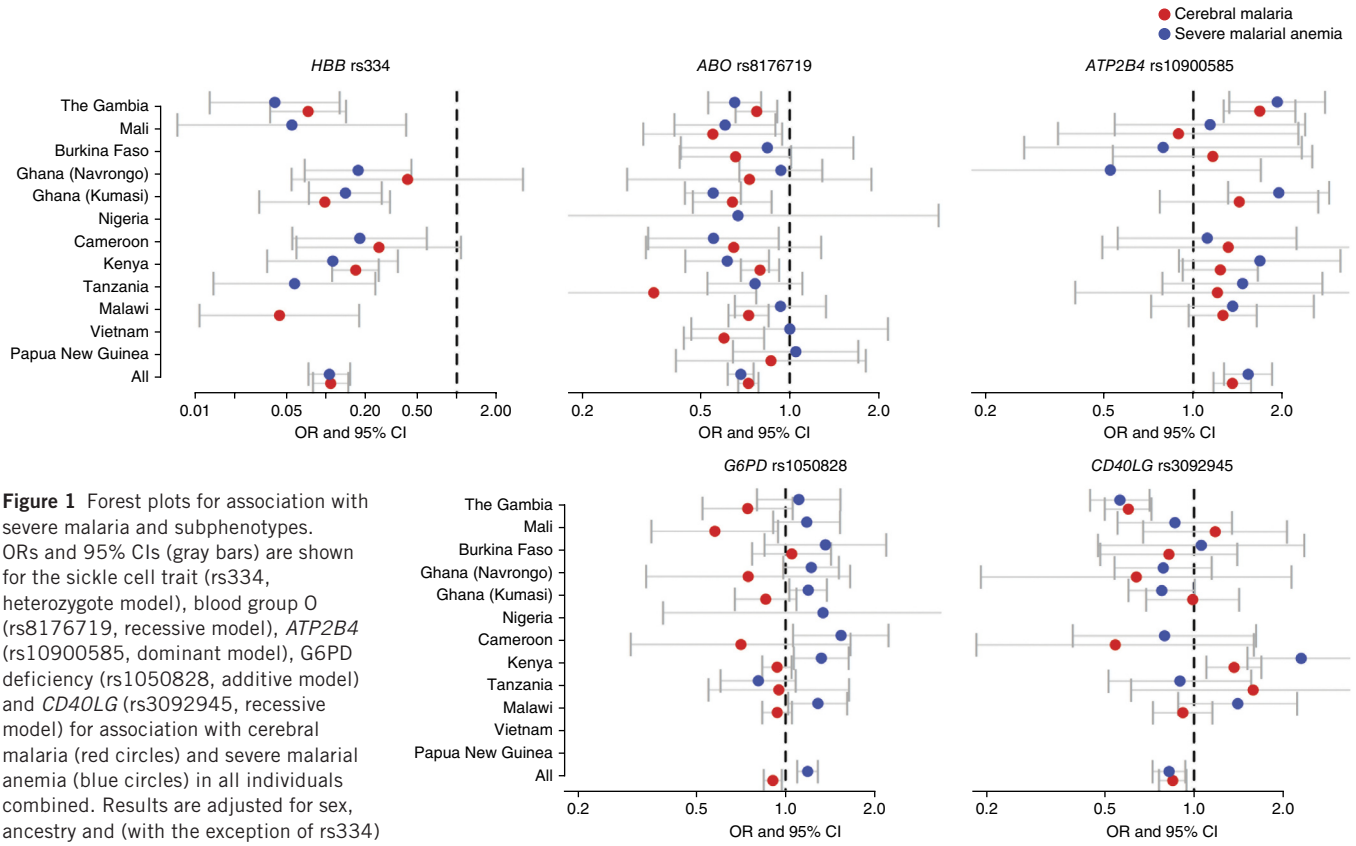


Figure 1 Forest plots for association with severe malaria and subphenotypes. ORs and 95% CIs (gray bars) are shown for the sickle cell trait (rs334, heterozygote model), blood group O (rs8176719, recessive model), *ATP2B4* (rs10900585, dominant model), G6PD deficiency (rs1050828, additive model) and *CD40LG* (rs3092945, recessive model) for association with cerebral malaria (red circles) and severe malarial anemia (blue circles) in all individuals combined. Results are adjusted for sex, ancestry and (with the exception of rs334) the sickle cell trait. Results are not presented when the sample size was too small (fewer than five cases or controls with the relevant genotype) or for locations where the derived allele was absent. Further details are available in **Supplementary Tables 11–19**. OR = 1, representing no effect, is highlighted by the vertical dashed lines.

ATP2B4. The *ATP2B4* gene, encoding a calcium transporter found in the plasma membrane of erythrocytes, has been identified by genome-wide association study (GWAS) as a malaria resistance locus[26]. We typed four SNPs in this gene that were found to be in LD; the derived alleles of rs10900585 and rs55868763 were associated with increased risk of severe malaria, whereas the derived alleles of rs4951074 and rs1541255 were associated with decreased risk (**Table 2** and **Supplementary Tables 8–10** and **18**). When aggregated across all African sites, individuals carrying at least one copy of the derived allele at rs10900585 had an OR of 1.32 for severe malaria ($P = 1.7 \times 10^{-9}$), whereas individuals homozygous for the derived allele at rs4951074 had an OR of 0.77 ($P = 7.6 \times 10^{-7}$). In both cases, the magnitude of the genetic effect was similar for cerebral malaria and severe malarial anemia (**Fig. 1**).

CD40LG. The *CD40LG* gene is a gene on the X chromosome encoding CD40 ligand that has previously been associated with severe malaria[27]. Homozygotes for the derived allele of a SNP in the 5′ UTR (rs3092945) showed reduced risk of severe malaria (OR = 0.85; $P = 1.1 \times 10^{-6}$), with a similar trend of protection in both males (OR = 0.90; $P = 0.01$) and females (OR = 0.78; $P = 8.9 \times 10^{-5}$) when the data were aggregated across sites (**Table 3**). However, when sites were analyzed individually, the results were strikingly different between sites: homozygotes for the derived allele showed significantly reduced risk of severe malaria in The Gambia (OR = 0.54; $P = 2.3 \times 10^{-22}$) but significantly increased risk in Kenya (OR = 1.42; $P = 7.8 \times 10^{-6}$) (**Supplementary Table 19**).

Other loci. None of the other loci tested here showed consistent evidence of association with severe malaria in the multicenter analysis with a significance of $P < 1 \times 10^{-4}$. All variants tested, some of which had weak associations that merit further investigation, are shown in **Supplementary Figure 3** and **Supplementary Tables 8–10**. At the *CD36* locus, heterozygotes for the codon variant rs201346212 tended to have reduced risk of severe malaria (OR = 0.67; $P = 4.2 \times 10^{-4}$). Other weak signals of association (P values in the range of 0.05 to 0.001) were observed for *CD36*, *IL1A* and *IRF1* with severe malaria overall, for *CR1* and *IL4* with cerebral malaria and for *IL20RA* with severe malarial anemia. Although it is clear from these data that many genetic associations reported in the literature might have been false positives, as has been observed for other common diseases[28], it is undoubtedly also the case that authentic genetic associations might be missed by multicenter studies if the effect is weak and there is heterogeneity of effect across different study sites.

Epistasis between significantly associated loci

Epistasis between malaria resistance loci has been reported in previous studies[29,30]. We therefore tested for pairwise interaction between all SNPs that showed significant association at the *HBB*, *ABO*, *G6PD*, *ATP2B4* and *CD40LG* loci (**Supplementary Fig. 4** and **Supplementary Table 20**). This analysis did not identify any strong evidence of interaction, but a marginally significant effect was observed between the *ATP2B4* locus (rs10900585) and the allele for HbC (rs33930165; $P = 1.3 \times 10^{-3}$), such that the ancestral allele of rs10900585, which was the minor allele in Africa, tended to reverse the protective effect of the HbC allele. This association warrants further investigation, as *ATP2B4* encodes the major erythrocyte calcium channel and intracellular calcium levels have been noted to affect the clinical phenotype of sickling disorders[31].

ARTICLES

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

The Wellcome Trust Case Control Consortium*

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

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

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

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

Samples and experimental analyses

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

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

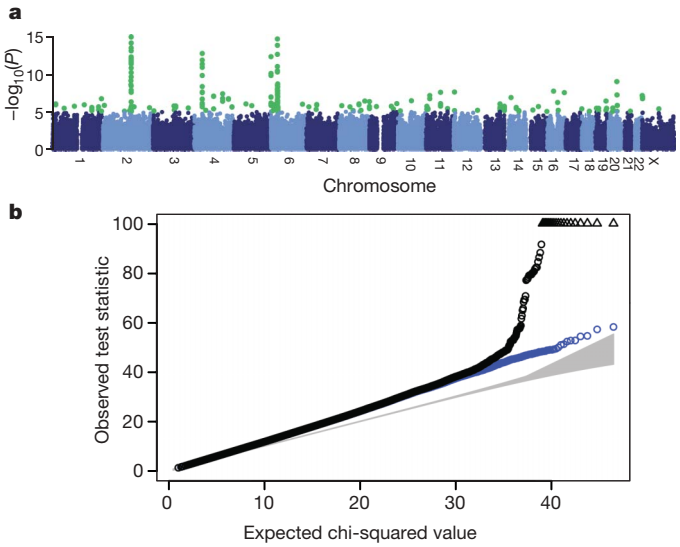


Figure 2 | Genome-wide picture of geographic variation. a, *P* values for the 11-d.f. test for difference in SNP allele frequencies between geographical regions, within the 9 collections. SNPs have been excluded using the project quality control filters described in Methods. Green dots indicate SNPs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

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

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

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

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

Disease association results

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

Table 1 | Highly differentiated SNPs

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

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