


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

View metadata, citation and similar papers at core.ac.uk

brought to you by  CORE

provided by Carolina Digital Repository

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 exomes¹ and whole genomes²⁻⁴ 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 differentiation⁶⁻⁸. 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 phase² 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.

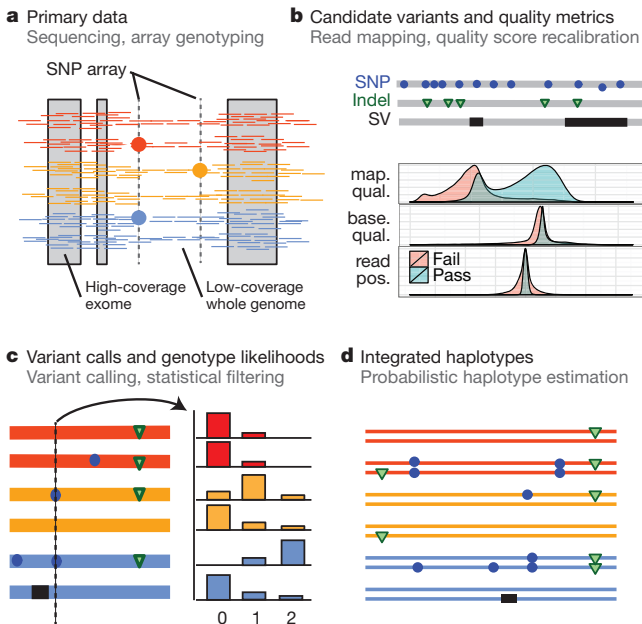
*Autosomal genes only.

†Compared with dbSNP release 135 (Oct 2011), excluding contribution from phase I 1000 Genomes Project (or equivalent data for large deletions).

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 groups². Primary data generated for each sample consist of low-coverage (average $5\times$) whole-genome and high-coverage (average $80\times$ 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 ($\leq 0.1\%$, 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 $\sim 50\%$, 98% and 99.7% of the SNPs with frequencies of $\sim 0.1\%$, 1.0% and 5.0%, respectively, in $\sim 2,500$ UK-sampled genomes (the Wellcome Trust-funded UK10K project), thus

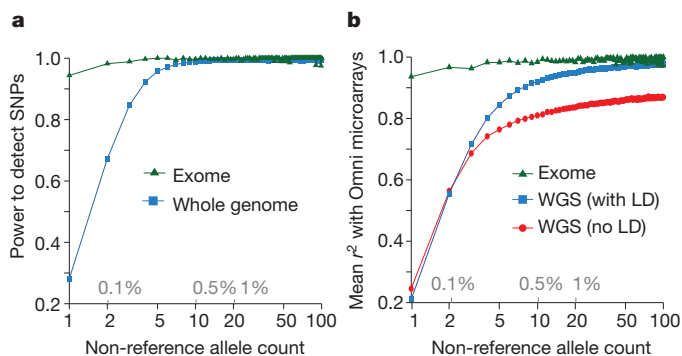


Figure 1 | Power and accuracy. **a**, Power to detect SNPs as a function of variant count (and proportion) across the entire set of samples, estimated by comparison to independent SNP array data in the exome (green) and whole genome (blue). **b**, Genotype accuracy compared with the same SNP array data 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.

meeting this goal. However, coverage may be lower for populations not closely related to those studied. For example, our resource includes only 23.7%, 76.9% and 99.3% of the SNPs with frequencies of $\sim 0.1\%$, 1.0% and 5.0%, respectively, in $\sim 2,000$ genomes sequenced in a study of the isolated population of Sardinia (the SardiNIA study).

Genetic variation within and between populations

The integrated data set provides a detailed view of variation across several populations (illustrated in Fig. 2a). Most common variants (94% of variants with frequency $\geq 5\%$ in Fig. 2a) were known before the current phase of the project and had their haplotype structure mapped through earlier projects^{2,9}. By contrast, only 62% of variants in the range 0.5–5% and 13% of variants with frequencies of $\leq 0.5\%$ had been described previously. For analysis, populations are grouped by the predominant component of ancestry: Europe (CEU) (see Fig. 2a for definitions of this and other populations), TSI, GBR, FIN and IBS), Africa (YRI, LWK and ASW), East Asia (CHB, JPT and CHS) and the Americas (MXL, CLM and PUR). Variants present at 10% and above across the entire sample are almost all found in all of the populations studied. By contrast, 17% of low-frequency variants in the range 0.5–5% were observed in a single ancestry group, and 53% of rare variants at 0.5% were observed in a single population (Fig. 2b). Within ancestry groups, common variants are weakly differentiated (most within-group estimates of Wright's fixation index (F_{ST}) are $< 1\%$; Supplementary Table 11), although below 0.5% frequency variants are up to twice as likely to be found within the same population compared with random samples from the ancestry group (Supplementary Fig. 6a). The degree of rare-variant differentiation varies between populations. For example, within Europe, the IBS and FIN populations carry excesses of rare variants (Supplementary Fig. 6b), which can arise through events such as recent bottlenecks¹⁰, 'clan' breeding structures¹¹ and admixture with diverged populations¹².

Some common variants show strong differentiation between populations within ancestry-based groups (Supplementary Table 12), many of which are likely to have been driven by local adaptation either directly or through hitchhiking. For example, the strongest differentiation between African populations is within an NRSF (neuron-restrictive silencer factor) transcription-factor peak (PANC1 cell line)¹³, upstream of *S78SIA1* (difference in derived allele frequency LWK – YRI of 0.475 at rs7960970), whose product is involved in ganglioside generation¹⁴. Overall, we find a range of 17–343 SNPs (fewest = CEU – GBR, most = FIN – TSI) showing a difference in frequency of at least 0.25 between pairs of populations within an ancestry group.

The derived allele frequency distribution shows substantial divergence between populations below a frequency of 40% (Fig. 2c), such that individuals from populations with substantial African ancestry (YRI, LWK and ASW) carry up to three times as many low-frequency variants (0.5–5% frequency) as those of European or East Asian origin, reflecting ancestral bottlenecks in non-African populations¹⁵. However, individuals from all populations show an enrichment of rare variants ($< 0.5\%$ frequency), reflecting recent explosive increases in population size and the effects of geographic differentiation^{6,16}. Compared with the expectations from a model of constant population size, individuals from all populations show a substantial excess of high-frequency-derived variants ($> 80\%$ frequency).

Because rare variants are typically recent, their patterns of sharing can reveal aspects of population history. Variants present twice across the entire sample (referred to as f_2 variants), typically the most recent of informative mutations, are found within the same population in 53% of cases (Fig. 3a). However, between-population sharing identifies recent historical connections. For example, if one of the individuals carrying an f_2 variant is from the Spanish population (IBS) and the other is not (referred to as IBS–X), the other individual is more likely to come from the Americas populations (48%, correcting for sample size) than from elsewhere in Europe (41%). Within the East Asian populations, CHS and CHB show stronger f_2 sharing to each other

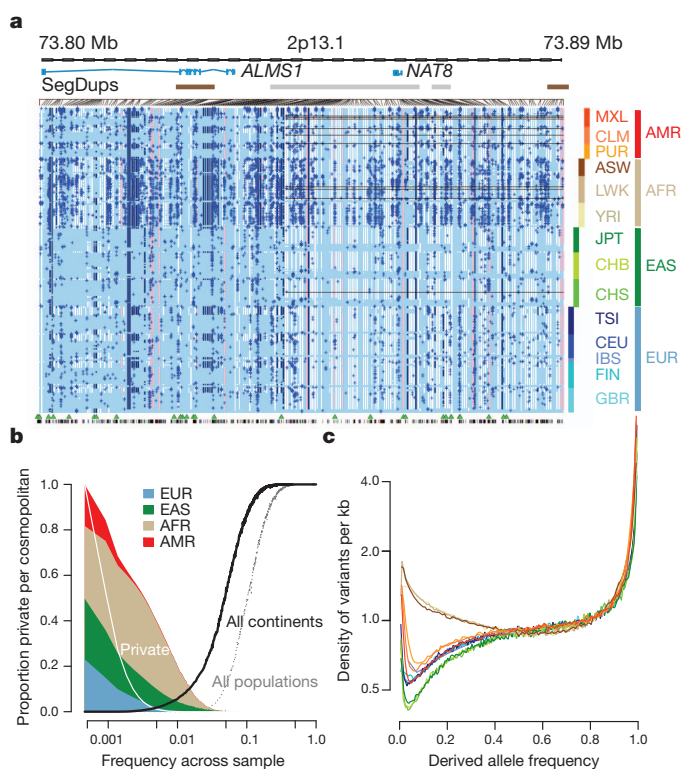


Figure 2 | The distribution of rare and common variants. **a**, Summary of inferred haplotypes across a 100-kb region of chromosome 2 spanning the genes *ALMS1* and *NAT8*, variation in which has been associated with kidney disease¹⁵. Each row represents an estimated haplotype, with the population of origin indicated on the right. Reference alleles are indicated by the light blue background. Variants (non-reference alleles) above 0.5% frequency are indicated by pink (typed on the high-density SNP array), white (previously known) and dark blue (not previously known). Low frequency variants ($< 0.5\%$) are indicated by blue crosses. Indels are indicated by green triangles and novel variants by dashes below. A large, low-frequency deletion (black line) spanning *NAT8* is present in some populations. Multiple structural haplotypes mediated by segmental duplications are present at this locus, including copy number gains, which were not genotyped for this study. Within each population, haplotypes are ordered by total variant count across the region. Population abbreviations: ASW, people with African ancestry in Southwest United States; CEU, Utah residents with ancestry from Northern and Western Europe; CHB, Han Chinese in Beijing, China; CHS, Han Chinese South, China; CLM, Colombians in Medellin, Colombia; FIN, Finnish in Finland; GBR, British from England and Scotland, UK; IBS, Iberian populations in Spain; LWK, Luhya in Webuye, Kenya; JPT, Japanese in Tokyo, Japan; MXL, people with Mexican ancestry in Los Angeles, California; PUR, Puerto Ricans in Puerto Rico; TSI, Toscani in Italia; YRI, Yoruba in Ibadan, Nigeria. Ancestry-based groups: AFR, African; AMR, Americas; EAS, East Asian; EUR, European. **b**, The fraction of variants identified across the project that are found in only one population (white line), are restricted to a single ancestry-based group (defined as in **a**, solid colour), are found in all groups (solid black line) and all populations (dotted black line). **c**, The density of the expected number of variants per kilobase carried by a genome drawn from each population, as a function of variant frequency (see Supplementary Information). Colours as in **a**. Under a model of constant population size, the expected density is constant across the frequency spectrum.

(58% and 53% of CHS–X and CHB–X variants, respectively) than either does to JPT, but JPT is closer to CHB than to CHS (44% versus 35% of JPT–X variants). Within African-ancestry populations, the ASW are closer to the YRI (42% of ASW–X f_2 variants) than the LWK (28%), in line with historical information¹⁷ and genetic evidence based on common SNPs¹⁸. Some sharing patterns are surprising; for example, 2.5% of the f_2 FIN–X variants are shared with YRI or LWK populations.

Independent evidence about variant age comes from the length of the shared haplotypes on which they are found. We find, as expected,

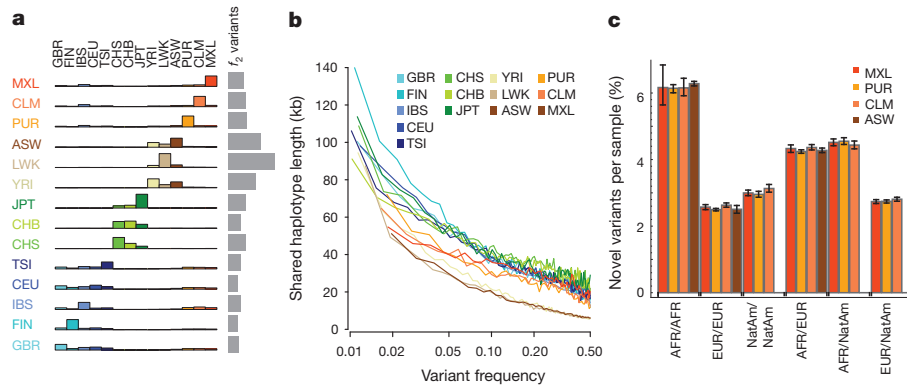


Figure 3 | Allele sharing within and between populations. **a**, Sharing of f_2 variants, those found exactly twice across the entire sample, within and between populations. Each row represents the distribution across populations for the origin of samples sharing an f_2 variant with the target population (indicated by the left-hand side). The grey bars represent the average number of f_2 variants carried by a randomly chosen genome in each population. **b**, Median length of haplotype identity (excluding cryptically related samples and singleton variants, and allowing for up to two genotype errors) between two

a negative correlation between variant frequency and the median length of shared haplotypes, such that chromosomes carrying variants at 1% frequency share haplotypes of 100–150 kb (typically 0.08–0.13 cM; Fig. 3b and Supplementary Fig. 7a), although the distribution is highly skewed and 2–5% of haplotypes around the rarest SNPs extend over 1 megabase (Mb) (Supplementary Fig. 7b, c). Haplotype phasing and genotype calling errors will limit the ability to detect long shared haplotypes, and the observed lengths are a factor of 2–3 times shorter than predicted by models that allow for recent explosive growth⁶ (Supplementary Fig. 7a). Nevertheless, the haplotype length for variants shared within and between populations is informative about relative allele age. Within populations and between populations in which there is recent shared ancestry (for example, through admixture and within continents), f_2 variants typically lie on long shared haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry, f_2 variants are present on very short haplotypes, for example, an average of 11 kb for FIN – YRI f_2 variants (median between ancestry groups excluding admixture is 15 kb), and are therefore likely to reflect recurrent mutations and chance ancient coalescent events.

To analyse populations with substantial historical admixture, statistical methods were applied to each individual to infer regions of the genome with different ancestries. Populations and individuals vary substantially in admixture proportions. For example, the MXL population contains the greatest proportion of Native American ancestry (47% on average compared with 24% in CLM and 13% in PUR), but the proportion varies from 3% to 92% between individuals (Supplementary Fig. 9a). Rates of variant discovery, the ratio of non-synonymous to synonymous variation and the proportion of variants that are new vary systematically between regions with different ancestries. Regions of Native American ancestry show less variation, but a higher fraction of the variants discovered are novel (3.0% of variants per sample; Fig. 3c) compared with regions of European ancestry (2.6%). Regions of African ancestry show the highest rates of novelty (6.2%) and heterozygosity (Supplementary Fig. 9b, c).

The functional spectrum of human variation

The phase I data enable us to compare, for different genomic features and variant types, the effects of purifying selection on evolutionary conservation¹⁹, the allele frequency distribution and the level of differentiation between populations. At the most highly conserved coding sites, 85% of non-synonymous variants and more than 90% of stop-gain and splice-disrupting variants are below 0.5% in frequency,

chromosomes that share variants of a given frequency in each population. Estimates are from 200 randomly sampled regions of 1 Mb each and up to 15 pairs of individuals for each variant. **c**, The average proportion of variants that are new (compared with the pilot phase of the project) among those found in regions inferred to have different ancestries within ASW, PUR, CLM and MXL populations. Error bars represent 95% bootstrap confidence intervals. NatAm, Native American.

compared with 65% of synonymous variants (Fig. 4a). In general, the rare variant excess tracks the level of evolutionary conservation for variants of most functional consequence, but varies systematically between types (for example, for a given level of conservation enhancer variants have a higher rare variant excess than variants in transcription-factor motifs). However, stop-gain variants and, to a lesser extent, splice-site disrupting changes, show increased rare-variant excess whatever the conservation of the base in which they occur, as such mutations can be highly deleterious whatever the level of sequence conservation. Interestingly, the least conserved splice-disrupting variants show similar rare-variant loads to synonymous and non-coding regions, suggesting that these alternative transcripts are under very weak selective constraint. Sites at which variants are observed are typically less conserved than average (for example, sites with non-synonymous variants are, on average, as conserved as third codon positions; Supplementary Fig. 10).

A simple way of estimating the segregating load arising from rare, deleterious mutations across a set of genes comes from comparing the

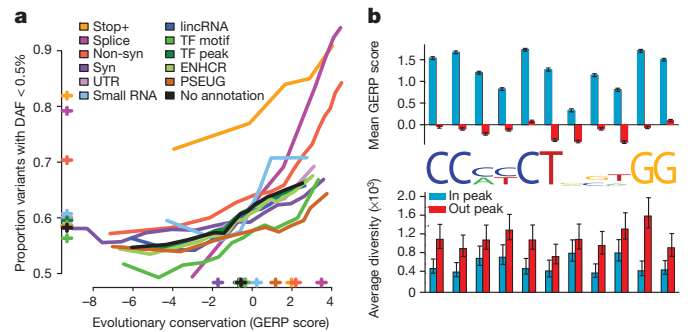


Figure 4 | Purifying selection within and between populations. **a**, The relationship between evolutionary conservation (measured by GERP score¹⁹) and rare variant proportion (fraction of all variants with derived allele frequency (DAF) < 0.5%) for variants occurring in different functional elements and with different coding consequences. Crosses indicate the average GERP score at variant sites (x axis) and the proportion of rare variants (y axis) in each category. ENHCR, enhancer; lincRNA, large intergenic non-coding RNA; non-syn, non-synonymous; PSEUG, pseudogene; syn, synonymous; TF, transcription factor. **b**, Levels of evolutionary conservation (mean GERP score, top) and genetic diversity (per-nucleotide pairwise differences, bottom) for sequences matching the CTCF-binding motif within CTCF-binding peaks, as identified experimentally by ChIP-seq in the ENCODE project¹³ (blue) and in a matched set of motifs outside peaks (red). The logo plot shows the distribution of identified motifs within peaks. Error bars represent ± 2 s.e.m.

ratios of non-synonymous to synonymous variants in different frequency ranges. The non-synonymous to synonymous ratio among rare (<0.5%) variants is typically in the range 1–2, and among common variants in the range 0.5–1.5, suggesting that 25–50% of rare non-synonymous variants are deleterious. However, the segregating rare load among gene groups in KEGG pathways²⁰ varies substantially (Supplementary Fig. 11a and Supplementary Table 13). Certain groups (for example, those involving extracellular matrix (ECM)–receptor interactions, DNA replication and the pentose phosphate pathway) show a substantial excess of rare coding mutations, which is only weakly correlated with the average degree of evolutionary conservation. Pathways and processes showing an excess of rare functional variants vary between continents (Supplementary Fig. 11b). Moreover, the excess of rare non-synonymous variants is typically higher in populations of European and East Asian ancestry (for example, the ECM–receptor interaction pathway load is strongest in European populations). Other groups of genes (such as those associated with allograft rejection) have a high non-synonymous to synonymous ratio in common variants, potentially indicating the effects of positive selection.

Genome-wide data provide important insights into the rates of functional polymorphism in the non-coding genome. For example, we consider motifs matching the consensus for the transcriptional repressor CTCF, which has a well-characterized and highly conserved binding motif²¹. Within CTCF-binding peaks experimentally defined by chromatin-immunoprecipitation sequencing (ChIP-seq), the average levels of conservation within the motif are comparable to third codon positions, whereas there is no conservation outside peaks (Fig. 4b). Within peaks, levels of genetic diversity are typically reduced 25–75%, depending on the position in the motif (Fig. 4b). Unexpectedly, the reduction in diversity at some degenerate positions, for example, at position 8 in the motif, is as great as that at non-degenerate positions, suggesting that motif degeneracy may not have a simple relationship with functional importance. Variants within peaks show a weak but consistent excess of rare variation (proportion with frequency <0.5% is 61% within peaks compared with 58% outside peaks; Supplementary Fig. 12), supporting the hypothesis that regulatory sequences contain substantial amounts of weakly deleterious variation.

Purifying selection can also affect population differentiation if its strength and efficacy vary among populations. Although the magnitude of the effect is weak, non-synonymous variants consistently show

greater levels of population differentiation than synonymous variants, for variants of frequencies of less than 10% (Supplementary Fig. 13).

Uses of 1000 Genomes Project data in medical genetics

Data from the 1000 Genomes Project are widely used to screen variants discovered in exome data from individuals with genetic disorders²² and in cancer genome projects²³. The enhanced catalogue presented here improves the power of such screening. Moreover, it provides a ‘null expectation’ for the number of rare, low-frequency and common variants with different functional consequences typically found in randomly sampled individuals from different populations.

Estimates of the overall numbers of variants with different sequence consequences are comparable to previous values^{1,20–22} (Supplementary Table 14). However, only a fraction of these are likely to be functionally relevant. A more accurate picture of the number of functional variants is given by the number of variants segregating at conserved positions (here defined as sites with a genomic evolutionary rate profiling (GERP)¹⁹ conservation score of >2), or where the function (for example, stop-gain variants) is strong and independent of conservation (Table 2). We find that individuals typically carry more than 2,500 non-synonymous variants at conserved positions, 20–40 variants identified as damaging²⁴ at conserved sites and about 150 loss-of-function (LOF) variants (stop-gains, frameshift indels in coding sequence and disruptions to essential splice sites). However, most of these are common (>5%) or low-frequency (0.5–5%), such that the numbers of rare (<0.5%) variants in these categories (which might be considered as pathological candidates) are much lower; 130–400 non-synonymous variants per individual, 10–20 LOF variants, 2–5 damaging mutations, and 1–2 variants identified previously from cancer genome sequencing²⁵. By comparison with synonymous variants, we can estimate the excess of rare variants; those mutations that are sufficiently deleterious that they will never reach high frequency. We estimate that individuals carry an excess of 76–190 rare deleterious non-synonymous variants and up to 20 LOF and disease-associated variants. Interestingly, the overall excess of low-frequency variants is similar to that of rare variants (Table 2). Because many variants contributing to disease risk are likely to be segregating at low frequency, we recommend that variant frequency be considered when using the resource to identify pathological candidates.

The combination of variation data with information about regulatory function¹³ can potentially improve the power to detect pathological

Table 2 | Per-individual variant load at conserved sites

Variant type	Number of derived variant sites per individual			Excess rare deleterious	Excess low-frequency deleterious
	Derived allele frequency across sample				
	<0.5%	0.5–5%	>5%		
All sites	30–150 K	120–680 K	3.6–3.9 M	ND	ND
Synonymous*	29–120	82–420	1.3–1.4 K	ND	ND
Non-synonymous*	130–400	240–910	2.3–2.7 K	76–190†	77–130†
Stop-gain*	3.9–10	5.3–19	24–28	3.4–7.5‡	3.8–11‡
Stop-loss	1.0–1.2	1.0–1.9	2.1–2.8	0.81–1.1†	0.80–1.0†
HGMD-DM*	2.5–5.1	4.8–17	11–18	1.6–4.7‡	3.8–12‡
COSMIC*	1.3–2.0	1.8–5.1	5.2–10	0.93–1.6‡	1.3–2.0‡
Indel frameshift	1.0–1.3	11–24	60–66	ND§	3.2–11‡
Indel non-frameshift	2.1–2.3	9.5–24	67–71	ND§	0–0.73‡
Splice site donor	1.7–3.6	2.4–7.2	2.6–5.2	1.6–3.3‡	3.1–6.2‡
Splice site acceptor	1.5–2.9	1.5–4.0	2.1–4.6	1.4–2.6‡	1.2–3.3‡
UTR*	120–430	300–1,400	3.5–4.0 K	0–350‡	0–1.2 K‡
Non-coding RNA*	3.9–17	14–70	180–200	0.62–2.6‡	3.4–13‡
Motif gain in TF peak*	4.7–14	23–59	170–180	0–2.6‡	3.8–15‡
Motif loss in TF peak*	18–69	71–300	580–650	7.7–22‡	37–110‡
Other conserved*	2.0–9.9 K	7.1–39 K	120–130 K	ND	ND
Total conserved	2.3–11 K	7.7–42 K	130–150 K	150–510	250–1.3 K

Only sites in which ancestral state can be assigned with high confidence are reported. The ranges reported are across populations. COSMIC, Catalogue of Somatic Mutations in Cancer; HGMD-DM, Human Gene Mutation Database (HGMD) disease-causing mutations; TF, transcription factor; ND, not determined.

* Sites with GERP >2

† Using synonymous sites as a baseline.

‡ Using ‘other conserved’ as a baseline.

§ Rare indels were filtered in phase I.

non-coding variants. We find that individuals typically contain several thousand variants (and several hundred rare variants) in conserved (GERP conservation score >2) untranslated regions (UTR), non-coding RNAs and transcription-factor-binding motifs (Table 2). Within experimentally defined transcription-factor-binding sites, individuals carry 700–900 conserved motif losses (for the transcription factors analysed, see Supplementary Information), of which 18–69 are rare ($<0.5\%$) and show strong evidence for being selected against. Motif gains are rarer (~ 200 per individual at conserved sites), but they also show evidence for an excess of rare variants compared with conserved sites with no functional annotation (Table 2). Many of these changes are likely to have weak, slightly deleterious effects on gene regulation and function.

A second major use of the 1000 Genomes Project data in medical genetics is imputing genotypes in existing genome-wide association studies (GWAS)²⁶. For common variants, the accuracy of using the phase I data to impute genotypes at sites not on the original GWAS SNP array is typically 90–95% in non-African and approximately 90% in African-ancestry genomes (Fig. 5a and Supplementary Fig. 14a), which is comparable to the accuracy achieved with high-quality benchmark haplotypes (Supplementary Fig. 14b). Imputation accuracy is similar for intergenic SNPs, exome SNPs, indels and large deletions (Supplementary Fig. 14c), despite the different amounts of information about such variants and accuracy of genotypes. For low-frequency variants (1–5%), imputed genotypes have between 60% and 90% accuracy in all populations, including those with admixed ancestry (also comparable to the accuracy from trio-phased haplotypes; Supplementary Fig. 14b).

Imputation has two primary uses: fine-mapping existing association signals and detecting new associations. GWAS have had only a few examples of successful fine-mapping to single causal variants^{27,28}, often because of extensive haplotype structure within regions of association^{29,30}. We find that, in Europeans, each previously reported GWAS signal³¹ is, on average, in linkage disequilibrium ($r^2 \geq 0.5$) with 56 variants: 51.5 SNPs and 4.5 indels. In 19% of cases at least one of these variants changes the coding sequence of a nearby gene (compared with 12% in control variants matched for frequency, distance to nearest gene and ascertainment in GWAS arrays) and in 65% of cases

at least one of these is at a site with GERP >2 (68% in matched controls). The size of the associated region is typically <200 kb in length (Fig. 5b). Our observations suggest that trans-ethnic fine-mapping experiments are likely to be especially valuable: among the 56 variants that are in strong linkage disequilibrium with a typical GWAS signal, approximately 15 show strong disequilibrium across our four continental groupings (Supplementary Table 15). Our current resource increases the number of variants in linkage disequilibrium with each GWAS signal by 25% compared with the pilot phase of the project and by greater than twofold compared with the HapMap resource.

Discussion

The success of exome sequencing in Mendelian disease genetics³² and the discovery of rare and low-frequency disease-associated variants in genes associated with complex diseases^{27,33,34} strongly support the hypothesis that, in addition to factors such as epistasis^{35,36} and gene-environment interactions³⁷, many other genetic risk factors of substantial effect size remain to be discovered through studies of rare variation. The data generated by the 1000 Genomes Project not only aid the interpretation of all genetic-association studies, but also provide lessons on how best to design and analyse sequencing-based studies of disease.

The use and cost-effectiveness of collecting several data types (low-coverage whole-genome sequence, targeted exome data, SNP genotype data) for finding variants and reconstructing haplotypes are demonstrated here. Exome capture provides private and rare variants that are missed by low-coverage data (approximately 60% of the singleton variants in the sample were detected only from exome data compared with 5% detected only from low-coverage data; Supplementary Fig. 15). However, whole-genome data enable characterization of functional non-coding variation and accurate haplotype estimation, which are essential for the analysis of *cis*-effects around genes, such as those arising from variation in upstream regulatory regions³⁸. There are also benefits from integrating SNP array data, for example, to improve genotype estimation³⁹ and to aid haplotype estimation where array data have been collected on additional family members. In principle, any sources of genotype information (for example, from array CGH) could be integrated using the statistical methods developed here.

Major methodological advances in phase I, including improved methods for detecting and genotyping variants⁴⁰, statistical and machine-learning methods for evaluating the quality of candidate variant calls, modelling of genotype likelihoods and performing statistical haplotype integration⁴¹, have generated a high-quality resource. However, regions of low sequence complexity, satellite regions, large repeats and many large-scale structural variants, including copy-number polymorphisms, segmental duplications and inversions (which constitute most of the ‘inaccessible genome’), continue to present a major challenge for short-read technologies. Some issues are likely to be improved by methodological developments such as better modelling of read-level errors, integrating *de novo* assembly^{42,43} and combining multiple sources of information to aid genotyping of structurally diverse regions^{40,44}. Importantly, even subtle differences in data type, data processing or algorithms may lead to systematic differences in false-positive and false-negative error modes between samples. Such differences complicate efforts to compare genotypes between sequencing studies. Moreover, analyses that naively combine variant calls and genotypes across heterogeneous data sets are vulnerable to artefact. Analyses across multiple data sets must therefore either process them in standard ways or use meta-analysis approaches that combine association statistics (but not raw data) across studies.

Finally, the analysis of low-frequency variation demonstrates both the pervasive effects of purifying selection at functionally relevant sites in the genome and how this can interact with population history to lead to substantial local differentiation, even when standard metrics of structure such as F_{ST} are very small. The effect arises primarily

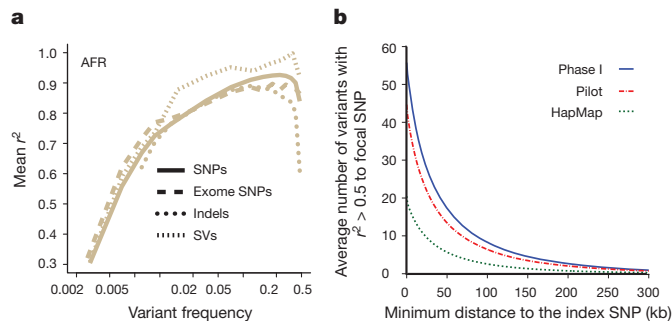


Figure 5 | Implications of phase I 1000 Genomes Project data for GWAS.

a, Accuracy of imputation of genome-wide SNPs, exome SNPs and indels (using sites on the Illumina 1 M array) into ten individuals of African ancestry (three LWK, four Maaai from Kinyawa, Kenya (MKK), two YRI), sequenced to high coverage by an independent technology³. Only indels in regions of high sequence complexity with frequency $>1\%$ are analysed. Deletion imputation accuracy estimated by comparison to array data⁴⁶ (note that this is for a different set of individuals, although with a similar ancestry, but included on the same plot for clarity). Accuracy measured by squared Pearson correlation coefficient between imputed and true dosage across all sites in a frequency range estimated from the 1000 Genomes data. Lines represent whole-genome SNPs (solid), exome SNPs (long dashes), short indels (dotted) and large deletions (short dashes). SV, structural variants. **b**, The average number of variants in linkage disequilibrium ($r^2 > 0.5$ among EUR) to focal SNPs identified in GWAS⁴⁷ as a function of distance from the index SNP. Lines indicate the number of HapMap (green), pilot (red) and phase I (blue) variants.

because rare variants tend to be recent and thus geographically restricted^{6–8}. The implication is that the interpretation of rare variants in individuals with a particular disease should be within the context of the local (either geographic or ancestry-based) genetic background. Moreover, it argues for the value of continuing to sequence individuals from diverse populations to characterize the spectrum of human genetic variation and support disease studies across diverse groups. A further 1,500 individuals from 12 new populations, including at least 15 high-depth trios, will form the final phase of this project.

METHODS SUMMARY

All details concerning sample collection, data generation, processing and analysis can be found in the Supplementary Information. Supplementary Fig. 1 summarizes the process and indicates where relevant details can be found.

Received 4 July; accepted 1 October 2012.

- Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
- The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
- Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
- Mills, R. E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
- Marth, G. T. *et al.* The functional spectrum of low-frequency coding variation. *Genome Biol.* **12**, R84 (2011).
- Nelson, M. R. *et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**, 100–104 (2012).
- Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nature Genet.* **44**, 243–246 (2012).
- Gravel, S. *et al.* Demographic history and rare allele sharing among human populations. *Proc. Natl Acad. Sci. USA* **108**, 11983–11988 (2011).
- The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
- Salmela, E. *et al.* Genome-wide analysis of single nucleotide polymorphisms uncovers population structure in Northern Europe. *PLoS ONE* **3**, e3519 (2008).
- Lupski, J. R., Belmont, J. W., Boerwinkle, E. & Gibbs, R. A. Clan genomics and the complex architecture of human disease. *Cell* **147**, 32–43 (2011).
- Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
- ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9**, e1001046 (2011).
- Sasaki, K. *et al.* Expression cloning of a novel Galβ(1–3/1–4)GlcNAc α2,3-sialyltransferase using lectin resistance selection. *J. Biol. Chem.* **268**, 22782–22787 (1993).
- Marth, G. *et al.* Sequence variations in the public human genome data reflect a bottlenecked population history. *Proc. Natl Acad. Sci. USA* **100**, 376–381 (2003).
- Keinan, A. & Clark, A. G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336**, 740–743 (2012).
- Hall, G. M. *Slavery and African Ethnicities in the Americas: Restoring the Links* (Univ. North Carolina Press, 2005).
- Bryc, K. *et al.* Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc. Natl Acad. Sci. USA* **107**, 786–791 (2010).
- Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6**, e1001025 (2010).
- Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109–D114 (2012).
- Kim, T. H. *et al.* Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128**, 1231–1245 (2007).
- Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Rev. Genet.* **12**, 745–755 (2011).
- Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
- Stenson, P. D. *et al.* The Human Gene Mutation Database: 2008 update. *Genome Med.* **1**, 13 (2009).
- Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39**, D945–D950 (2011).
- Howie, B., Marchini, J. & Stephens, M. Genome imputation with thousands of genomes. *G3 (Bethesda)* **1**, 457–470 (2011).
- Sanna, S. *et al.* Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. *PLoS Genet.* **7**, e1002198 (2011).
- Gregory, A. P., Dendrou, C. A., Bell, J., McVean, G. & Fugger, L. TNF receptor 1 genetic risk mirrors outcome of anti-TNF therapy in multiple sclerosis. *Nature* **488**, 508–511 (2012).
- Hassanein, M. T. *et al.* Fine mapping of the association with obesity at the *FTO* locus in African-derived populations. *Hum. Mol. Genet.* **19**, 2907–2916 (2010).
- Maller, J. R. The Wellcome Trust Case Control Consortium. Fine mapping of 14 loci identified through genome-wide association analyses. *Nature Genet.* (in the press).

- Hindorf, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106**, 9362–9367 (2009).
- Bamshad, M. J. *et al.* The Centers for Mendelian Genomics: A new large-scale initiative to identify the genes underlying rare Mendelian conditions. *Am. J. Med. Genet. A.* (2012).
- Momozawa, Y. *et al.* Resequencing of positional candidates identifies low frequency *IL23R* coding variants protecting against inflammatory bowel disease. *Nature Genet.* **43**, 43–47 (2011).
- Raychaudhuri, S. *et al.* A rare penetrant mutation in *CFH* confers high risk of age-related macular degeneration. *Nature Genet.* **43**, 1232–1236 (2011).
- Strange, A. *et al.* A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between *HLA-C* and *ERAP1*. *Nature Genet.* **42**, 985–990 (2010).
- Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl Acad. Sci. USA* **109**, 1193–1198 (2012).
- Thomas, D. Gene-environment-wide association studies: emerging approaches. *Nature Rev. Genet.* **11**, 259–272 (2010).
- Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).
- Flannick, J. *et al.* Efficiency and power as a function of sequence coverage, SNP array density, and imputation. *PLoS Comput. Biol.* **8**, e1002604 (2012).
- Handaker, R. E., Korn, J. M., Nemes, J. & McCarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nature Genet.* **43**, 269–276 (2011).
- Li, Y., Sidore, C., Kang, H. M., Boehnke, M. & Abecasis, G. R. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.* **21**, 940–951 (2011).
- Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. *De novo* assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genet.* **44**, 226–232 (2012).
- Simpson, J. T. & Durbin, R. Efficient construction of an assembly string graph using the FM-index. *Bioinformatics* **26**, i367–i373 (2010).
- Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).
- Chambers, J. C. *et al.* Genetic loci influencing kidney function and chronic kidney disease. *Nature Genet.* **42**, 373–375 (2010).
- Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
- Hindorf, L. A. *et al.* A Catalog of Published Genome-Wide Association Studies. Available at <http://www.genome.gov/gwastudies> (accessed, September 2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank many people who contributed to this project: A. Naranjo, M. V. Parra and C. Duque for help with the collection of the Colombian samples; N. Kälin and F. Laplace for discussions; A. Schlattl and T. Zichner for assistance in managing data sets; E. Appelbaum, H. Arbery, E. Birney, S. Bumpstead, J. Camarata, J. Carey, G. Cochrane, M. DaSilva, S. Dökel, E. Drury, C. Duque, K. Gyaltsen, P. Jokinen, B. Lenz, S. Lewis, D. Lu, A. Naranjo, S. Ott, I. Padioulea, M. V. Parra, N. Patterson, A. Price, L. Sadzewicz, S. Schrinner, N. Sengamalai, J. Sullivan, F. Ta, Y. Vaydylevich, O. Venn, K. Watkins and A. Yurovsky for assistance, discussion and advice. We thank the people who generously contributed their samples, from these populations: Yoruba in Ibadan, Nigeria; the Han Chinese in Beijing, China; the Japanese in Tokyo, Japan; the Utah CEPH community; the Luhya in Webuye, Kenya; people with African ancestry in the Southwest United States; the Toscani in Italy; people with Mexican ancestry in Los Angeles, California; the Southern Han Chinese in China; the British in England and Scotland; the Finnish in Finland; the Iberian Populations in Spain; the Colombians in Medellín, Colombia; and the Puerto Ricans in Puerto Rico. This research was supported in part by Wellcome Trust grants WT098051 to R.M.D., M.E.H. and C.T.S.; WT090532/Z/09/Z, WT085475/Z/08/Z and WT095552/Z/11/Z to P.D.; WT086084/Z/08/Z and WT090532/Z/09/Z to G.A.M.; WT089250/Z/09/Z to I.M.; WT085532AIA to P.F.; Medical Research Council grant G0900747(91070) to G.A.M.; British Heart Foundation grant RG/09/12/28096 to C.A.A.; the National Basic Research Program of China (973 program no. 2011CB809201, 2011CB809202 and 2011CB809203); the Chinese 863 program (2012AA02A201); the National Natural Science Foundation of China (30890032, 31161130357); the Shenzhen Key Laboratory of Transomics Biotechnologies (CXB201108250096A); the Shenzhen Municipal Government of China (grants ZYC200903240080A and ZYC201105170397A); Guangdong Innovative Research Team Program (no. 2009010016); BMBF grant 01GS08201 to H.Le.; BMBF grant 015428A to R.H.; the Max Planck Society; Swiss National Science Foundation 31003A_130342 to E.T.D.; Swiss National Science Foundation NCCR 'Frontiers in Genetics' grant to E.T.D.; Louis Jeantet Foundation grant to E.T.D.; Biotechnology and Biological Sciences Research Council (BBSRC) grant BB/I021213/1 to A.R.-L.; German Research Foundation (Emmy Noether Fellowship KO 4037/1-1) to J.O.K.; Netherlands Organization for Scientific Research VENI grant 639.021.125 to K.Y.; Beatriu de Pinós Program grants 2006BP-A 10144 and 2009BP-B 00274 to M.V.; Israeli Science Foundation grant 04514831 to E.H.; Genome Québec and the Ministry of Economic Development, Innovation and Trade grant PSR-SIIRI-195 to P.A.w.; National Institutes of Health (NIH) grants U01HG5214, RC2HG5581 and R01MH84698 to G.R.A.; R01HG4719 and R01HG3698 to G.T.M.; RC2HG5552 and U01HG6513 to G.R.A. and G.T.M.; R01HG4960 and R01HG5701 to B.L.B.; U01HG5715 to C.D.B. and A.G.C.; T32GM8283 to D.C.I.; U01HG5208 to M.J.D.; U01HG6569 to M.A.D.; R01HG2898 and R01CA166661 to S.E.D.; U01HG5209, U01HG5725 and P41HG4221 to C.Le.; P01HG4120 to E.E.E.; U01HG5728 to Yu.F.; U54HG3273 and U01HG5211 to R.A.G.;

R01HL95045 to S.B.G.; U41HG4568 to S.J.K.; P41HG2371 to W.J.K.; ES015794, AI077439, HL088133 and HL078885 to E.G.B.; RC2HL102925 to S.B.G. and D.M.A.; R01GM59290 to L.B.J. and M.A.B.; U54HG3067 to E.S.L. and S.B.G.; T15LM7033 to B.K.M.; T32HL94284 to J.L.R.-F.; DP2OD6514 and BAA-NIAID-DAIT-NIHAI2009061 to P.C.S.; T32GM7748 to X.S.; U54HG3079 to R.K.W.; UL1R024131 to R.D.H.; HHSN268201100040C to the Coriell Institute for Medical Research; a Sandler Foundation award and an American Asthma Foundation award to E.G.B.; an IBM Open Collaborative Research Program award to Y.B.; an A.G. Leventis Foundation scholarship to D.K.X.; a Wolfson Royal Society Merit Award to P.Do.; a Howard Hughes Medical Institute International Fellowship award to P.H.S.; a grant from T. and V. Stanley to S.C.Y.; and a Mary Beryl Patch Turnbull Scholar Program award to K.C.B. E.H. is a faculty fellow of the Edmond J. Safra Bioinformatics program at Tel-Aviv University. E.E.E. and D.H. are investigators of the Howard Hughes Medical Institute. M.V.G. is a long-term fellow of EMBO.

Author Contributions Details of author contributions can be found in the author list.

Author Information All primary data, alignments, individual call sets, consensus call sets, integrated haplotypes with genotype likelihoods and supporting data including details of validation are available from the project website (<http://www.1000genomes.org>). Variant and haplotypes for specific genomic regions and specific samples can be viewed and downloaded through the project browser (<http://browser.1000genomes.org/>). Common project variants with no known medical impact have been compiled by dbSNP for filtering (http://www.ncbi.nlm.nih.gov/variation/docs/human_variation_vcf/). The authors declare competing financial interests: details are available in the online version of the paper. Reprints and permissions information is available at www.nature.com/reprints. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to G.A.M. (mcvan@well.ox.ac.uk). This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported licence. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-sa/3.0/>

The 1000 Genomes Consortium (Participants are arranged by project role, then by institution alphabetically, and finally alphabetically within institutions except for Principal Investigators and Project Leaders, as indicated.)

Corresponding author Gil A. McVean^{1,2}

Steering committee David M. Altshuler^{3,4,5} (Co-Chair), Richard M. Durbin⁶ (Co-Chair), Gonçalo R. Abecasis⁷, David R. Bentley⁸, Aravinda Chakravarti⁹, Andrew G. Clark¹⁰, Peter Donnelly^{1,2}, Evan E. Eichler¹¹, Paul Flicek¹², Stacey B. Gabriel³, Richard A. Gibbs¹³, Eric D. Green¹⁴, Matthew E. Hurles⁵, Bartha M. Knoppers¹⁵, Jan O. Korbel¹⁶, Eric S. Lander³, Charles Lee¹⁷, Hans Lehrach^{18,19}, Elaine R. Mardis²⁰, Gabor T. Marth²¹, Gil A. McVean^{1,2}, Deborah A. Nickerson²², Jeanette P. Schmidt²³, Stephen T. Sherry²⁴, Jun Wang^{25,26,27}, Richard K. Wilson²⁰

Production group: Baylor College of Medicine Richard A. Gibbs¹³ (Principal Investigator), Huyen Dinh¹³, Christie Kovar¹³, Sandra Lee¹³, Lora Lewis¹³, Donna Muzny¹³, Jeff Reid¹³, Min Wang¹³; **BGI-Shenzhen** Jun Wang^{25,26,27} (Principal Investigator), Xiaodong Fang²⁵, Xiaosen Guo²⁵, Min Jian²⁵, Hui Jiang²⁵, Xin Jin²⁵, Guoqing Li²⁵, Jingxiang Li²⁵, Yingrui Li²⁵, Zhuo Li²⁵, Xiao Liu²⁵, Yao Lu²⁵, Xuedi Ma²⁵, Zhe Su²⁵, Shuaishuai Tai²⁵, Meifang Tang²⁵, Bo Wang²⁵, Guangbiao Wang²⁵, Honglong Wu²⁵, Renhua Wu²⁵, Ye Yin²⁵, Wenwei Zhang²⁵, Jiao Zhao²⁵, Meiru Zhao²⁵, Xiaole Zheng²⁵, Yan Zhou²⁵; **Broad Institute of MIT and Harvard** Eric S. Lander³ (Principal Investigator), David M. Altshuler^{3,4,5}, Stacey B. Gabriel³ (Co-Chair), Namrata Gupta³; **European Bioinformatics Institute** Paul Flicek¹² (Principal Investigator), Laura Clarke¹², Rasko Leinonen¹², Richard E. Smith¹², Xiangqun Zheng-Bradley¹²; **llumina** David R. Bentley⁸ (Principal Investigator), Russell Grocock⁸, Sean Humphray⁸, Terena James⁸, Zoya Kingsbury⁸; **Max Planck Institute for Molecular Genetics** Hans Lehrach^{18,19} (Principal Investigator), Ralf Sudbrak¹⁸ (Project Leader), Marcus W. Albrecht²⁸, Vyacheslav S. Amstislavskiy¹⁸, Tatiana A. Borodina²⁸, Matthias Lienhard¹⁸, Florian Mertes¹⁸, Marc Sultan¹⁸, Bernd Timmermann¹⁸, Marie-Laure Yaspo¹⁸; **US National Institutes of Health** Stephen T. Sherry²⁴ (Principal Investigator); **University of Oxford** Gil A. McVean^{1,2} (Principal Investigator); **Washington University in St Louis** Elaine R. Mardis²⁰ (Co-Principal Investigator) (Co-Chair), Richard K. Wilson²⁰ (Co-Principal Investigator), Lucinda Fulton²⁰, Robert Fulton²⁰, George M. Weinstock²⁰; **Wellcome Trust Sanger Institute** Richard M. Durbin⁶ (Principal Investigator), Senduran Balasubramaniam⁶, John Burton⁶, Petr Danecek⁶, Thomas M. Keane⁶, Anja Kolb-Kokocinski⁶, Shane McCarthy⁶, James Stalker⁶, Michael Quail⁶

Analysis group: Affymetrix Jeanette P. Schmidt²³ (Principal Investigator), Christopher J. Davies²³, Jeremy Gollub²³, Teresa Webster²³, Brant Wong²³, Yiping Zhan²³; **Albert Einstein College of Medicine** Adam Auton²⁹ (Principal Investigator); **Baylor College of Medicine** Richard A. Gibbs¹³ (Principal Investigator), Fuli Yu¹³ (Project Leader), Matthew Bainbridge¹³, Danny Challis¹³, Uday S. Evani¹³, James Lu¹³, Donna Muzny¹³, Uma Nagaswamy¹³, Jeff Reid¹³, Aniko Sobo¹³, Yi Wang¹³, Jin Yu¹³; **BGI-Shenzhen** Jun Wang^{25,26,27} (Principal Investigator), Lachlan J. M. Coin²⁵, Lin Fang²⁵, Xiaosen Guo²⁵, Xin Jin²⁵, Guoqing Li²⁵, Qibin Li²⁵, Yingrui Li²⁵, Zhenyu Li²⁵, Haoxiang Lin²⁵, Binghang Liu²⁵, Ruibang Luo²⁵, Nan Qin²⁵, Haojing Shao²⁵, Bingqiang Wang²⁵, YONGLONG Xie²⁵, Chen Ye²⁵, Chang Yu²⁵, Fan Zhang²⁵, Hancheng Zheng²⁵, Hongmei Zhu²⁵; **Boston College** Gabor T. Marth²¹ (Principal Investigator), Erik P. Garrison²¹, Deniz Kural²¹, Wan-Ping Lee²¹, Wen Fung Leong²¹, Alistair N. Ward²¹, Jiantao Wu²¹, Mengyao

Zhang²¹; **Brigham and Women's Hospital** Charles Lee¹⁷ (Principal Investigator), Lauren Griffin¹⁷, Chih-Heng Hsieh¹⁷, Ryan E. Mills^{17,30}, Xinghua Shi¹⁷, Marc von Grotthuss¹⁷, Chengsheng Zhang¹⁷; **Broad Institute of MIT and Harvard** Mark J. Daly³ (Principal Investigator), Mark A. DePristo³ (Project Leader), David M. Altshuler^{3,4,5}, Eric Banks³, Gaurav Bhatia³, Mauricio O. Carneiro³, Guillermo del Angel³, Stacey B. Gabriel³, Giulio Genovese³, Namrata Gupta³, Robert E. Handsaker^{3,5}, Chris Hartl³, Eric S. Lander³, Steven A. McCarroll³, James C. Nemes³, Ryan E. Poplin³, Stephen F. Schaffner³, Khalid Shakir³; **Cold Spring Harbor Laboratory** Seungtae C. Yoon³¹ (Principal Investigator), Jayon Lihm³¹, Vladimir Makarov³²; **Dankook University** Hanjun Jin³³ (Principal Investigator), Wook Kim³⁴, Ki Cheol Kim³⁴; **European Molecular Biology Laboratory** Jan O. Korbel¹⁶ (Principal Investigator), Tobias Rausch¹⁶; **European Bioinformatics Institute** Paul Flicek¹² (Principal Investigator), Kathryn Beal¹², Laura Clarke¹², Fiona Cunningham¹², Javier Herrero¹², William M. McLaren¹², Graham R. S. Ritchie¹², Richard E. Smith¹², Xiangqun Zheng-Bradley¹²; **Cornell University** Andrew G. Clark¹⁰ (Principal Investigator), Srikanth Gottipati³⁵, Alon Keinan¹⁰, Juan L. Rodriguez-Flores¹⁰; **Harvard University** Pardis C. Sabeti^{3,36} (Principal Investigator), Sharon R. Grossman^{3,36}, Shervin Tabrizi^{3,36}, Ridhi Tariyal^{3,36}; **Human Gene Mutation Database** David N. Cooper³⁷ (Principal Investigator), Edward V. Ball³⁷, Peter D. Stenson³⁷; **llumina** David R. Bentley⁸ (Principal Investigator), Bret Barnes³⁸, Markus Bauer³⁸, R. Keira Cheetham³⁸, Tony Cox³⁸, Michael Eberle³⁸, Sean Humphray³⁸, Scott Kahn³⁸, Lisa Murray³⁸, John Peden³⁸, Richard Shaw³⁸; **Leiden University Medical Center** Kai Ye³⁹ (Principal Investigator); **Louisiana State University** Mark A. Batzer⁴⁰ (Principal Investigator), Miriam K. Konke⁴⁰, Jerilyn A. Walker⁴⁰; **Massachusetts General Hospital** Daniel G. MacArthur⁴¹ (Principal Investigator), Monkol Lek⁴¹; **Max Planck Institute for Molecular Genetics** Ralf Sudbrak¹⁸ (Project Leader), Vyacheslav S. Amstislavskiy¹⁸, Ralf Herwig¹⁸; **Pennsylvania State University** Mark D. Shriver⁴² (Principal Investigator); **Stanford University** Carlos D. Bustamante⁴³ (Principal Investigator), Jake K. Byrnes⁴³, Francisco M. De La Vega¹⁰, Simon Gravel⁴³, Eimear E. Kenny⁴³, Jeffrey M. Kidd⁴³, Phil Lacroute⁴³, Brian K. Maples⁴³, Andres Moreno-Estrada⁴³, Fouad Zakharia⁴³; **Tel-Aviv University** Eran Halperin^{45,46,47} (Principal Investigator), Yael Baran⁴⁵; **Translational Genomics Research Institute** David W. Craig⁴⁸ (Principal Investigator), Alexis Christoforides⁴⁸, Nils Homer⁴⁹, Tyler Izatt⁴⁸, Ahmet A. Kurdoglu⁴⁸, Shripad A. Sinari⁴⁸, Kevin Squire⁵⁰; **US National Institutes of Health** Stephen T. Sherry²⁴ (Principal Investigator), Chunlin Xiao²⁴; **University of California, San Diego** Jonathan Sebat^{51,52} (Principal Investigator), Vineet Bafna⁵³, Kenny Ye⁵⁴; **University of California, San Francisco** Esteban G. Burchard⁵⁵ (Principal Investigator), Ryan D. Hernandez⁵⁵ (Principal Investigator), Christopher R. Gignoux⁵⁵; **University of California, Santa Cruz** David Haussler^{56,57} (Principal Investigator), Sol J. Katzman⁵⁶, W. James Kent⁵⁶; **University of Chicago** Bryan Howie⁵⁸; **University College London** Andres Ruiz-Linares⁵⁹ (Principal Investigator); **University of Geneva** Emmanouil T. Dermizakis^{60,61,62} (Principal Investigator), Tuuli Lappalainen^{60,61,62}; **University of Maryland School of Medicine** Scott E. Devine⁶³ (Principal Investigator), Xinyue Liu⁶³, Ankit Morok⁶³, Luke J. Tallon⁶³; **University of Medicine and Dentistry of New Jersey** Jeffrey A. Rosenfeld^{64,65} (Principal Investigator), Leslie P. Michelson⁶⁴; **University of Michigan** Gonçalo R. Abecasis⁷ (Principal Investigator) (Co-Chair), Hyun Min Kang⁷ (Project Leader), Paul Anderson⁷, Andrea Angius⁶⁶, Abigail Bigham⁶⁷, Tom Blackwell⁶⁷, Fabio Busonero^{7,66,68}, Francesco Cucca^{66,68}, Christian Fuchsberger⁶⁷, Chris Jones⁶⁹, Goo Jun⁷, Yun Li⁷⁰, Robert Lyons⁷¹, Andrea Maschio^{7,66,68}, Eleonora Porcu^{7,66,68}, Fred Reinier⁶⁹, Serena Sanna⁶⁶, David Schlessinger⁷², Carlo Sidore^{7,66,68}, Adrian Tan⁷, Mary Kate Trost⁷; **University of Montreal** Philip Awadalla⁷³ (Principal Investigator), Alan Hodgkinson⁷³; **University of Oxford** Gerton Lunter¹ (Principal Investigator), Gil A. McVean^{1,2} (Principal Investigator) (Co-Chair), Jonathan L. Marchini^{1,2} (Principal Investigator), Simon Myers^{1,2} (Principal Investigator), Claire Churchhouse², Olivier Delaneau², Anjali Gupta-Hinch¹, Zamin Iqbal¹, Iain Mathieson¹, Andy Rimmer¹, Dionysia K. Xifara^{1,2}; **University of Puerto Rico** Taras K. Oleksyk⁷⁴ (Principal Investigator); **University of Texas Health Sciences Center at Houston** Yunxin Fu⁷⁵ (Principal Investigator), Xiaoming Liu⁷⁵, Momiao Xiong⁷⁵; **University of Utah** Lynn Jorde⁷⁶ (Principal Investigator), David Witherspoon⁷⁶, Jinchuan Xing⁷⁷; **University of Washington** Evan E. Eichler¹¹ (Principal Investigator), Brian L. Browning⁷⁸ (Principal Investigator), Can Alkan^{22,79}, Iman Hajirasouliha⁸⁰, Fereydoon Hormozdian²², Arthur Ko²², Peter H. Sudmant²²; **Washington University in St Louis** Elaine R. Mardis²⁰ (Co-Principal Investigator), Ken Chen⁸¹, Asif Chinwalla²⁰, Li Ding²⁰, David Dooling²⁰, Daniel C. Koboldt²⁰, Michael D. McLellan²⁰, John W. Wallis²⁰, Michael C. Wendt²⁰, Qunyan Zhang²⁰; **Wellcome Trust Sanger Institute** Richard M. Durbin⁶ (Principal Investigator), Matthew E. Hurles⁵ (Principal Investigator), Chris Tyler-Smith⁶ (Principal Investigator), Cornelis A. Albers⁸², Qasim Ayub⁶, Senduran Balasubramaniam⁶, Yuan Chen⁶, Alison J. Coffey⁶, Vincenza Colonna^{6,83}, Petr Danecek⁶, Ni Huang⁶, Luke Jostins⁶, Thomas M. Keane⁶, Heng Li^{3,6}, Shane McCarthy⁶, Aylwyn Scally⁶, James Stalker⁶, Klaudia Walter⁶, Yali Xue⁶, Yujun Zhang⁶; **Yale University** Mark B. Gerstein^{84,85,86} (Principal Investigator), Alexej Abyzov^{84,86}, Suganthi Balasubramanian⁸⁶, Jieming Chen⁸⁴, Declan Clarke⁸⁷, Yao Fu⁸⁴, Lukas Habegger⁸⁴, Arif O. Harmancl⁸⁴, Mike Jin⁸⁶, Ekta Khurana⁸⁶, Xinmeng Jasmine Mu⁸⁴, Cristina Sisu⁸⁴

Structural variation group: BGI-Shenzhen Yingrui Li²⁵, Ruibang Luo²⁵, Hongmei Zhu²⁵; **Brigham and Women's Hospital** Charles Lee¹⁷ (Principal Investigator) (Co-Chair), Lauren Griffin¹⁷, Chih-Heng Hsieh¹⁷, Ryan E. Mills^{17,30}, Xinghua Shi¹⁷, Marc von Grotthuss¹⁷, Chengsheng Zhang¹⁷; **Boston College** Gabor T. Marth²¹ (Principal Investigator), Erik P. Garrison²¹, Deniz Kural²¹, Wan-Ping Lee²¹, Alistair N. Ward²¹, Jiantao Wu²¹, Mengyao Zhang²¹; **Broad Institute of MIT and Harvard** Steven A. McCarroll³ (Project Leader), David M. Altshuler^{3,4,5}, Eric Banks³, Guillermo del Angel³, Giulio Genovese³, Robert E. Handsaker^{3,5}, Chris Hartl³, James C. Nemes³, Khalid Shakir³; **Cold Spring Harbor Laboratory** Seungtae C. Yoon³¹ (Principal Investigator), Jayon Lihm³¹, Vladimir Makarov³²; **Cornell University** Jeremiah Degenhardt¹⁰; **European Bioinformatics Institute** Paul Flicek¹² (Principal Investigator), Laura Clarke¹², Richard E. Smith¹², Xiangqun Zheng-Bradley¹²;

European Molecular Biology Laboratory Jan O. Korbel¹⁶ (Principal Investigator) (Co-Chair), Tobias Rausch¹⁰, Adrian M. Stütz¹⁶, **llumina** David R. Bentley⁸ (Principal Investigator), Bret Barnes³⁸, R. Keira Cheetham⁸, Michael Eberle⁸, Sean Humphray⁸, Scott Kahn³⁸, Lisa Murray⁸, Richard Shaw⁸, **Leiden University Medical Center** Kai Ye³⁹ (Principal Investigator); **Louisiana State University** Mark A. Batzer⁴⁰ (Principal Investigator), Miriam K. Konkel⁴⁰, Jeryllyn A. Walker⁴⁰; **Stanford University** Phil Lacroute⁴³; **Translational Genomics Research Institute** David W. Craig⁴⁸ (Principal Investigator), Nils Homer⁴⁹; **US National Institutes of Health** Deanna Church²⁴, Chunlin Xiao²⁴; **University of California, San Diego** Jonathan Sebat^{51,52} (Principal Investigator), Vineet Bafna⁵³, Jacob J. Michaelson⁸⁸, Kenny Ye⁵⁴; **University of Maryland School of Medicine** Scott E. Devine⁶³ (Principal Investigator), Xinyue Liu⁶³, Ankit Maroo⁶³, Luke J. Tallon⁶³; **University of Oxford** Gerton Lunter¹ (Principal Investigator), Gil A. McVean^{1,2} (Principal Investigator), Zamin Iqbal¹; **University of Utah** David Witherspoon⁷⁶, Jinchuan Xing⁷⁷; **University of Washington** Evan E. Eichler¹¹ (Principal Investigator) (Co-Chair), Can Alkan^{22,79}, Iman Hajirasouliha⁸⁰, Fereydoon Hormozdian²², Arthur Ko²², Peter H. Sudmant²²; **Washington University in St Louis** Ken Chen⁸¹, Asif Chinwalla²⁰, Li Ding²⁰, Michael D. McLellan²⁰, John W. Wallis²⁰; **Wellcome Trust Sanger Institute** Matthew E. Hurles⁶ (Principal Investigator) (Co-Chair), Ben Blackburne⁶, Heng Li^{3,6}, Sarah J. Lindsay⁶, Zemin Ning⁶, Aylwyn Scally⁶, Klaudia Walter⁶, Yujun Zhang⁶; **Yale University** Mark B. Gerstein^{84,85,86} (Principal Investigator), Alexej Abyzov^{84,86}, Jieming Chen⁸⁴, Declan Clarke⁸⁷, Ekta Khurana⁸⁶, Ximeng Jasmine Mu⁸⁴, Cristina Sisu⁸⁴

Exome group: Baylor College of Medicine Richard A. Gibbs¹³ (Principal Investigator) (Co-Chair), Fuli Yu¹³ (Project Leader), Matthew Bainbridge¹³, Danny Challis¹³, Uday S. Evans¹³, Christie Kovar¹³, Lora Lewis¹³, James Lu¹³, Donna Muzny¹³, Uma Nagaswamy¹³, Jeff Reid¹³, Aniko Sabo¹³, Jin Yu¹³; **BGI-Shenzhen** Xiaosen Guo²⁵, Yingrui Li²⁵, Renhua Wu²⁵; **Boston College** Gabor T. Marth²¹ (Principal Investigator) (Co-Chair), Erik P. Garrison²¹, Wen Fung Leong²¹, Alistair N. Ward²¹; **Broad Institute of MIT and Harvard** Guillermo del Angel³, Mark A. DePristo³, Stacey B. Gabriel³, Namrata Gupta³, Chris Hartl³, Ryan E. Poplin³; **Cornell University** Andrew G. Clark¹⁰ (Principal Investigator), Juan L. Rodriguez-Flores¹⁰; **European Bioinformatics Institute** Paul Flicek¹² (Principal Investigator), Laura Clarke¹², Richard E. Smith¹², Xiangqun Zheng-Bradley¹²; **Massachusetts General Hospital** Daniel G. MacArthur⁴¹ (Principal Investigator); **Stanford University** Carlos D. Bustamante⁴³ (Principal Investigator), Simon Gravel⁴³; **Translational Genomics Research Institute** David W. Craig⁴⁸ (Principal Investigator), Alexis Christoforides⁴⁸, Nils Homer⁴⁹, Tyler Izatt⁴⁸; **US National Institutes of Health** Stephen T. Sherry²⁴ (Principal Investigator), Chunlin Xiao²⁴; **University of Geneva** Emmanouil T. Dermitzakis^{60,61,62} (Principal Investigator); **University of Michigan** Gonçalo R. Abecasis⁷ (Principal Investigator), Hyun Min Kang⁷; **University of Oxford** Gil A. McVean^{1,2} (Principal Investigator); **Washington University in St Louis** Elaine R. Mardis²⁰ (Principal Investigator), David Dooling²⁰, Lucinda Fulton²⁰, Robert Fulton²⁰, Daniel C. Koboldt²⁰; **Wellcome Trust Sanger Institute** Richard M. Durbin⁶ (Principal Investigator), Senduran Balasubramaniam⁶, Thomas M. Keane⁶, Shane McCarthy⁶, James Stalker⁶; **Yale University** Mark B. Gerstein^{84,85,86} (Principal Investigator), Suganthi Balasubramaniam⁸⁶, Lukas Habegger⁸⁴

Functional interpretation group: Boston College Erik P. Garrison²¹; **Baylor College of Medicine** Richard A. Gibbs¹³ (Principal Investigator), Matthew Bainbridge¹³, Donna Muzny¹³, Fuli Yu¹³, Jin Yu¹³; **Broad Institute of MIT and Harvard** Guillermo del Angel³, Robert E. Handsaker^{3,5}; **Cold Spring Harbor Laboratory** Vladimir Makarov³²; **Cornell University** Juan L. Rodriguez-Flores¹⁰; **Dankook University** Hanjun Jin³³ (Principal Investigator), Wook Kim³⁴, Ki Cheol Kim³⁴; **European Bioinformatics Institute** Paul Flicek¹² (Principal Investigator), Kathryn Beal¹², Laura Clarke¹², Fiona Cunningham¹², Javier Herrero¹², William M. McLaren¹², Graham R. S. Ritchie¹², Xiangqun Zheng-Bradley¹²; **Harvard University** Sherwin Tabrizi^{3,36}; **Massachusetts General Hospital** Daniel G. MacArthur⁴¹ (Principal Investigator), Monkol Lek⁴¹; **Stanford University** Carlos D. Bustamante⁴³ (Principal Investigator), Francisco M. De La Vega¹⁰; **Translational Genomics Research Institute** David W. Craig⁴⁸ (Principal Investigator), Ahmet A. Kurdoglu⁴⁸; **University of Geneva** Tuuli Lappalainen^{60,61,62}; **University of Medicine and Dentistry of New Jersey** Jeffrey A. Rosenfeld^{64,65} (Principal Investigator), Leslie P. Michelson^{64,65}; **University of Montréal** Philip Awadalla⁷³ (Principal Investigator), Alan Hodgkinson⁷³; **University of Oxford** Gil A. McVean^{1,2} (Principal Investigator); **Washington University in St Louis** Ken Chen⁸¹; **Wellcome Trust Sanger Institute** Chris Tyler-Smith⁶ (Principal Investigator) (Co-Chair), Yuan Chen⁶, Vincenza Colonna^{6,83}, Adam Frankish⁶, Jennifer Harrow⁶, Yali Xue⁶; **Yale University** Mark B. Gerstein^{84,85,86} (Principal Investigator) (Co-Chair), Alexej Abyzov^{84,86}, Suganthi Balasubramaniam⁸⁶, Jieming Chen⁸⁴, Declan Clarke⁸⁷, Yao Fu⁸⁴, Arif O. Harmanci⁸⁴, Mike Jin⁸⁶, Ekta Khurana⁸⁶, Ximeng Jasmine Mu⁸⁴, Cristina Sisu⁸⁴

Data coordination centre group: Baylor College of Medicine Richard A. Gibbs¹³ (Principal Investigator), Gerald Fowler¹³, Walker Hale¹³, Divya Kaira¹³, Christie Kovar¹³, Donna Muzny¹³, Jeff Reid¹³; **BGI-Shenzhen** Jun Wang^{25,26,27} (Principal Investigator), Xiaosen Guo²⁵, Guoqing Li²⁵, Yingrui Li²⁵, Xiaole Zheng²⁵; **Broad Institute of MIT and Harvard** David M. Altshuler^{34,45}; **European Bioinformatics Institute** Paul Flicek¹² (Principal Investigator) (Co-Chair), Laura Clarke¹² (Project Leader), Jonathan Barker¹², Gavin Kelman¹², Eugene Kulesha¹², Rasko Leinonen¹², William M. McLaren¹², Rajesh Radhakrishnan¹², Asier Roa¹², Dmitriy Smirnov¹², Richard E. Smith¹², Ian Streeter¹², Iliana Toneva¹², Brendan Vaughan¹², Xiangqun Zheng-Bradley¹²; **llumina** David R. Bentley⁸ (Principal Investigator), Tony Cox⁸, Sean Humphray⁸, Scott Kahn³⁸; **Max Planck Institute for Molecular Genetics** Ralf Sudbrak¹⁸ (Project Leader), Marcus W. Albrecht²⁸, Matthias Lienhard¹⁸; **Translational Genomics Research Institute** David W. Craig⁴⁸ (Principal Investigator), Tyler Izatt⁴⁸, Ahmet A. Kurdoglu⁴⁸; **US National Institutes of Health** Stephen T. Sherry²⁴ (Principal Investigator) (Co-Chair), Victor Ananiev²⁴, Zinaida Belaia²⁴, Dimitriy Beloslyudtsev²⁴, Nathan Bouk²⁴, Chao Chen²⁴, Deanna Church²⁴, Robert Cohen²⁴, Charles Cook²⁴, John Garner²⁴, Timothy

Hefferon²⁴, Mikhail Kimelman²⁴, Chunlei Liu²⁴, John Lopez²⁴, Peter Meric²⁴, Chris O'Sullivan⁸⁹, Yuri Ostapchuk²⁴, Lon Phan²⁴, Sergiy Ponomarov²⁴, Valerie Schneider²⁴, Eugene Shekhtman²⁴, Karl Sirotkin²⁴, Douglas Slotta²⁴, Chunlin Xiao²⁴, Hua Zhang²⁴; **University of California, Santa Cruz** David Haussler^{56,57} (Principal Investigator); **University of Michigan** Gonçalo R. Abecasis⁷ (Principal Investigator); **University of Oxford** Gil A. McVean^{1,2} (Principal Investigator); **University of Washington** Can Alkan^{22,79}, Arthur Ko²²; **Washington University in St Louis** David Dooling²⁰; **Wellcome Trust Sanger Institute** Richard M. Durbin⁶ (Principal Investigator), Senduran Balasubramaniam⁶, Thomas M. Keane⁶, Shane McCarthy⁶, James Stalker⁶

Samples and ELSI group Aravinda Chakravarti⁹ (Co-Chair), Bartha M. Knoppers¹⁵ (Co-Chair), Gonçalo R. Abecasis⁷, Kathleen C. Barnes⁹⁰, Christine Beiswanger⁹¹, Esteban G. Burchard⁵⁵, Carlos D. Bustamante⁴³, Hongyu Cai²⁵, Hongzhi Cao²⁵, Richard M. Durbin⁶, Neda Gharani⁹¹, Richard A. Gibbs¹³, Christopher R. Gignoux⁵⁵, Simon Gravel⁴³, Brenna Henn⁴³, Danielle Jones³⁵, Lynn Jorde⁶, Jane S. Kaye⁹², Alon Keinan¹⁰, Alastair Kent⁹³, Angeliki Kerasidou⁹, Yingrui Li²⁵, Rasika Mathias⁹⁴, Gil A. McVean^{1,2}, Andres Moreno-Estrada⁴³, Pilar N. Ossorio^{95,96}, Michael Parker⁹⁷, David Reich⁵, Charles N. Rotimi⁹⁸, Charmaine D. Royal⁹⁹, Karla Sandoval⁴³, Yeyang Su²⁵, Ralf Sudbrak¹⁸, Zhongming Tian²⁵, Bernd Timmermann¹⁸, Sarah Tishkoff¹⁰⁰, Lorraine H. Toji⁹¹, Chris Tyler-Smith⁶, Marc Via¹⁰¹, Yuhong Wang²⁵, Huanming Yang²⁵, Ling Yang²⁵, Jiayong Zhu²⁵

Sample collection: British from England and Scotland (GBR) Walter Bodmer¹⁰²; **Colombians in Medellín, Colombia (CLM)** Gabriel Bedoya¹⁰³, Andres Ruiz-Linares⁵⁹; **Han Chinese South (CHS)** Cai Zhi Ming²⁵, Gao Yang¹⁰⁴, Chu Jia You¹⁰⁵; **Finnish in Finland (FIN)** Leena Peltonen[†]; **Iberian populations in Spain (IBS)** Andres Garcia-Montero¹⁰⁶, Alberto Orfao¹⁰⁷; **Puerto Ricans in Puerto Rico (PUR)** Julie Dutil¹⁰⁸, Juan C. Martinez-Cruzado⁷⁴, Taras K. Oleksyk⁷⁴

Scientific management Lisa D. Brooks¹⁰⁹, Adam L. Felsenfeld¹⁰⁹, Jean E. McEwen¹⁰⁹, Nicholas C. Clegg¹⁰⁹, Audrey Duncanson¹¹⁰, Michael Dunn¹¹⁰, Eric D. Green¹⁴, Mark S. Guyer¹⁰⁹, Jane L. Peterson¹⁰⁹

Writing group Gonçalo R. Abecasis⁷, Adam Auton²⁹, Lisa D. Brooks¹⁰⁹, Mark A. DePristo³, Richard M. Durbin⁶, Robert E. Handsaker^{3,5}, Hyun Min Kang⁷, Gabor T. Marth²¹, Gil A. McVean^{1,2}

¹Wellcome Trust Centre for Human Genetics, Oxford University, Oxford OX3 7BN, UK. ²Department of Statistics, Oxford University, Oxford OX1 3TG, UK. ³The Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. ⁴Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. ⁵Department of Genetics, Harvard Medical School, Cambridge, Massachusetts 02142, USA. ⁶Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SA, UK. ⁷Center for Statistical Genetics, Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA. ⁸Illumina United Kingdom, Chesterford Research Park, Little Chesterford, Near Saffron Walden, Essex CB10 1XL, UK. ⁹McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. ¹⁰Center for Comparative and Population Genomics, Cornell University, Ithaca, New York 14850, USA. ¹¹Department of Genome Sciences, University of Washington School of Medicine and Howard Hughes Medical Institute, Seattle, Washington 98195, USA. ¹²European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK. ¹³Baylor College of Medicine, Human Genome Sequencing Center, Houston, Texas 77030, USA. ¹⁴US National Institutes of Health, National Human Genome Research Institute, 31 Center Drive, Bethesda, Maryland 20892, USA. ¹⁵Centre of Genomics and Policy, McGill University, Montréal, Québec H3A 1A4, Canada. ¹⁶European Molecular Biology Laboratory, Genome Biology Research Unit, Meyerhofstraße 1, 69117 Heidelberg, Germany. ¹⁷Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. ¹⁸Max Planck Institute for Molecular Genetics, Ihnestraße 63-73, 14195 Berlin, Germany. ¹⁹Dahlem Centre for Genome Research and Medical Systems Biology, D-14195 Berlin-Dahlem, Germany. ²⁰The Genome Center, Washington University School of Medicine, St Louis, Missouri 63108, USA. ²¹Department of Biology, Boston College, Chestnut Hill, Massachusetts 02467, USA. ²²Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA. ²³Affymetrix, Inc., Santa Clara, California 95051, USA. ²⁴US National Institutes of Health, National Center for Biotechnology Information, 45 Center Drive, Bethesda, Maryland 20892, USA. ²⁵BGI-Shenzhen, Shenzhen 518083, China. ²⁶The Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, DK-2200 Copenhagen, Denmark. ²⁷Department of Biology, University of Copenhagen, DK-2100 Copenhagen, Denmark. ²⁸Alacris Theranostics GmbH, D-14195 Berlin-Dahlem, Germany. ²⁹Department of Genetics, Albert Einstein College of Medicine, Bronx, New York 10461, USA. ³⁰Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan 48109, USA. ³¹Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA. ³²Seaver Autism Center and Department of Psychiatry, Mount Sinai School of Medicine, New York, New York 10029, USA. ³³Department of Nanobiomedical Science, Dankook University, Cheonan 330-714, South Korea. ³⁴Department of Biological Sciences, Dankook University, Cheonan 330-714, South Korea. ³⁵Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853, USA. ³⁶Center for Systems Biology and Department Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA. ³⁷Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK. ³⁸Illumina, Inc., San Diego, California 92122, USA. ³⁹Molecular Epidemiology Section, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center 2333 ZA, The Netherlands. ⁴⁰Department of Biological Sciences, Louisiana State University, Baton

Rouge, Louisiana 70803, USA. ⁴¹Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. ⁴²Department of Anthropology, Penn State University, University Park, Pennsylvania 16802, USA. ⁴³Department of Genetics, Stanford University, Stanford, California 94305, USA. ⁴⁴Ancestry.com, San Francisco, California 94107, USA. ⁴⁵Blavatnik School of Computer Science, Tel-Aviv University, 69978 Tel Aviv, Israel. ⁴⁶Department of Microbiology, Tel-Aviv University, 69978 Tel Aviv, Israel. ⁴⁷International Computer Science Institute, Berkeley, California 94704, USA. ⁴⁸The Translational Genomics Research Institute, Phoenix, Arizona 85004, USA. ⁴⁹Life Technologies, Beverly, Massachusetts 01915, USA. ⁵⁰Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, California 90024, USA. ⁵¹Department of Psychiatry, University of California, San Diego, La Jolla, California 92093, USA. ⁵²Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, California 92093, USA. ⁵³Department of Computer Science, University of California, San Diego, La Jolla, California 92093, USA. ⁵⁴Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, New York 10461, USA. ⁵⁵Department of Bioengineering and Therapeutic Sciences and Medicine, University of California, San Francisco, California 94158, USA. ⁵⁶Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064, USA. ⁵⁷Howard Hughes Medical Institute, Santa Cruz, California 95064, USA. ⁵⁸Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA. ⁵⁹Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK. ⁶⁰Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland. ⁶¹Institute for Genetics and Genomics in Geneva (iGE3), University of Geneva, 1211 Geneva, Switzerland. ⁶²Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland. ⁶³Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA. ⁶⁴IST/High Performance and Research Computing, University of Medicine and Dentistry of New Jersey, Newark, New Jersey 07107, USA. ⁶⁵Department of Invertebrate Zoology, American Museum of Natural History, New York, New York 10024, USA. ⁶⁶Istituto di Ricerca Genetica e Biomedica, CNR, Monserrato, 09042 Cagliari, Italy. ⁶⁷Department of Anthropology, University of Michigan, Ann Arbor, Michigan 48109, USA. ⁶⁸Dipartimento di Scienze Biomediche, Università degli Studi di Sassari, 07100 Sassari, Italy. ⁶⁹Center for Advanced Studies, Research, and Development in Sardinia (CRS4), AGCT Program, Parco Scientifico e tecnologico della Sardegna, 09010 Pula, Italy. ⁷⁰Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. ⁷¹University of Michigan Sequencing Core, University of Michigan, Ann Arbor, Michigan 48109, USA. ⁷²National Institute on Aging, Laboratory of Genetics, Baltimore, Maryland 21224, USA. ⁷³Department of Pediatrics, University of Montréal, Ste. Justine Hospital Research Centre, Montréal, Québec H3T 1C5, Canada. ⁷⁴Department of Biology, University of Puerto Rico, Mayagüez, Puerto Rico 00680, USA. ⁷⁵The University of Texas Health Science Center at Houston, Houston, Texas 77030, USA. ⁷⁶Eccles Institute of Human Genetics, University of Utah School of Medicine, Salt Lake City, Utah 84112, USA. ⁷⁷Department of Genetics, Rutgers University, The State University of New Jersey, Piscataway, New Jersey 08854, USA. ⁷⁸Department of Medicine, Division of Medical Genetics, University of Washington, Seattle, Washington 98195, USA. ⁷⁹Department of

Computer Engineering, Bilkent University, TR-06800 Bilkent, Ankara, Turkey. ⁸⁰Department of Computer Science, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada. ⁸¹Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77230, USA. ⁸²Department of Haematology, University of Cambridge and National Health Service Blood and Transplant, Cambridge CB2 1TN, UK. ⁸³Institute of Genetics and Biophysics, National Research Council (CNR), 80125 Naples, Italy. ⁸⁴Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA. ⁸⁵Department of Computer Science, Yale University, New Haven, Connecticut 06520, USA. ⁸⁶Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA. ⁸⁷Department of Chemistry, Yale University, New Haven, Connecticut 06520, USA. ⁸⁸Beyster Center for Genomics of Psychiatric Diseases, University of California, San Diego, La Jolla, California 92093, USA. ⁸⁹US National Institutes of Health, National Human Genome Research Institute, 50 South Drive, Bethesda, Maryland 20892, USA. ⁹⁰Division of Allergy and Clinical Immunology, School of Medicine, Johns Hopkins University, Baltimore, Maryland 21205, USA. ⁹¹Coriell Institute for Medical Research, Camden, New Jersey 08103, USA. ⁹²Centre for Health, Law and Emerging Technologies, University of Oxford, Oxford OX3 7LF, UK. ⁹³Genetic Alliance, London N1 3QP, UK. ⁹⁴Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. ⁹⁵Department of Medical History and Bioethics, Morgridge Institute for Research, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA. ⁹⁶University of Wisconsin Law School, Madison, Wisconsin 53706, USA. ⁹⁷The Ethox Centre, Department of Public Health, University of Oxford, Old Road Campus, Oxford OX3 7LF, UK. ⁹⁸US National Institutes of Health, Center for Research on Genomics and Global Health, National Human Genome Research Institute, 12 South Drive, Bethesda, Maryland 20892, USA. ⁹⁹Institute for Genome Sciences and Policy, Duke University, Durham, North Carolina 27708, USA. ¹⁰⁰Department of Genetics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA. ¹⁰¹Department of Animal Biology, Unit of Anthropology, University of Barcelona, 08028 Barcelona, Spain. ¹⁰²Cancer and Immunogenetics Laboratory, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DS, UK. ¹⁰³Laboratory of Molecular Genetics, Institute of Biology, University of Antioquia, Medellín, Colombia. ¹⁰⁴Peking University Shenzhen Hospital, Shenzhen 518036, China. ¹⁰⁵Institute of Medical Biology, Chinese Academy of Medical Sciences and Peking Union Medical College, Kunming 650118, China. ¹⁰⁶Instituto de Biología Molecular y Celular del Cáncer, Centro de Investigación del Cáncer/IBMCC (CSIC-USAL), Institute of Biomedical Research of Salamanca (IBSAL) & Banco Nacional de ADN Carlos III, University of Salamanca, 37007 Salamanca, Spain. ¹⁰⁷Instituto de Biología Molecular y Celular del Cáncer, Centro de Investigación del Cáncer/IBMCC (CSIC-USAL), Institute of Biomedical Research of Salamanca (IBSAL) & Cytometry Service and Department of Medicine, 37007 University of Salamanca, Salamanca, Spain. ¹⁰⁸Ponce School of Medicine and Health Sciences, Ponce, Puerto Rico 00716, USA. ¹⁰⁹US National Institutes of Health, National Human Genome Research Institute, 5635 Fishers Lane, Bethesda, Maryland 20892, USA. ¹¹⁰Wellcome Trust, Gibbs Building, 215 Euston Road, London NW1 2BE, UK.

‡Deceased.