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 and whole genomes 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. Furthermore, because low-frequency variants tend to be recent in origin, they exhibit increased levels of population differentiation. 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 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

| 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.
† 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 FST typically < 1%) within broader geographical or ancestry-based groups. 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 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 biallelic 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 SNP 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 r² 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 ~0.1%, 1.0% and 5.0%, respectively, in ~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 ≥5% in Fig. 2a) were known before the current phase of the project and had their haplotype structure mapped through earlier projects. By contrast, only 62% of variants in the range 0.5–5% and 13% of variants with frequencies of ≤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 bottlenecks, ‘clan’ breeding structures and admixture with diverged populations.

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) upstream of ST8SIA1 (difference in derived allele frequency LWK − YRI of 0.475 at rs7609970), whose product is involved in ganglioside generation. 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 populations. 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. 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 (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 information and genetic evidence based on common SNPs. 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,
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 growth4 (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 haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8). By contrast, between populations with no recent shared ancestry haplotypes (median within ancestry group 103 kb; Supplementary Fig. 8).

### 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, 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 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.

---

**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 populations. Error bars represent 95% bootstrap confidence intervals. NatAm, Native American.

**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 project19 (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 pathways varies substantially (Supplementary Fig. 11a and Supplementary Table 13). Certain groups (for example, those involving extracellular matrix (ECM)–receptor interactions, DNA replication and the pentosine 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 motif. 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). Unsurprisingly, 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 disorders and in cancer genome projects. 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 (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) 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 damaging 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, 2–5 damaging mutations, and 1–2 variants identified previously from cancer genome sequencing.

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 function 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                                       | 1.0–19                   | 2.1–2.8                         | 0.8–1.1†                         | 0.8–1.0†                          |
| HGD-DOM*     | 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                          | 0.9–1.6†                         | 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                                   | 2.1–4.5                   | 2.9–4.6                         | 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                                     | 180–200                  | 0.62–2.6                         | 3.4–13‡                          |
| Motif gain in TF peak* | 4.7–14                                      | 23–59                     | 170–180                         | 0.2–6.2                          | 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                                     | 7.1–39 K                  | 120–130 K                       | ND                               | 7.0 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; HGD-DOM, 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). 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)\textsuperscript{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\textsuperscript{27,28}, often because of extensive haplotype structure within regions of association\textsuperscript{29,30}. We find that, in Europeans, each previously reported GWAS signal\textsuperscript{31} 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 (also comparable to the accuracy from trio-phased haplotypes; Supplementary Fig. 14b).

Discussion

The success of exome sequencing in Mendelian disease genetics\textsuperscript{32} and the discovery of rare and low-frequency disease-associated variants in genes associated with complex diseases\textsuperscript{27,33,34} strongly support the hypothesis that, in addition to factors such as epistasis\textsuperscript{35,36} and gene–environment interactions\textsuperscript{37}, 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 regions\textsuperscript{38}. There are also benefits from integrating SNP array data, for example, to improve genotype estimation\textsuperscript{19} 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 variants\textsuperscript{40}, statistical and machine-learning methods for evaluating the quality of candidate variant calls, modelling of genotype likelihoods and performing statistical haplotype integration\textsuperscript{41}, 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\textsuperscript{42,43} and combining multiple sources of information to aid genotyping of structurally diverse regions\textsuperscript{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](https://example.com/image.png)
because rare variants tend to be recent and thus geographically restricted. 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.**

1. Tennen, J. A. et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
2. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
3. Marth, G. T. et al. Human genome sequencing using reversible terminator chemistry. *Science* **327**, 71–75 (2009).
4. Mills, R. E. et al. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
5. Salmela, E. et al. Population-scale sequencing. *Nature Genet.* **44**, 243–246 (2012).
6. Marth, G. T. & Novembre, J. High-depth trio sequencing: implications for design of complex trait association studies. *Nature Genet.* **39**, 437–440 (2012).
7. Novembre, J. et al. Mapping copy number variation by population-scale genome sequencing. *Nature Genet.* **44**, 243–246 (2012).
8. Davydov, E. V. et al. Integrated genomic analyses of ovarian carcinoma. *Proc. Natl Acad. Sci. USA* **109**, 11933–11938 (2012).
9. Tishkoff, S. A. et al. The human genome sequence and the future of human biology. *Science* **337**, 944–950 (2012).
10. Ensembl Genome Browser. www.ensembl.org (accessed, September 2012).
11. Flannick, J. et al. Efficiency and power as a function of sequence coverage, SNP array density, and imputation. *PLOS Comput. Biol.* **8**, e1002604 (2012).
12. Schork, N. J. et al. A user’s guide to the encyclopedia of DNA elements. *Nat. Genet.* **43**, 43–47 (2011).
13. Maller, J., The Wellcome Trust Case Control Consortium. Fine mapping of the association with obesity at the FTO locus in African-derived populations. *Hum. Mol. Genet.* **19**, 2907–2916 (2010).
14. Maller, J., The Wellcome Trust Case Control Consortium. Fine mapping of 14 loci identified through genome-wide association analyses. *Nature Genet.* (in the press).

**Acknowledgements** We thank many people who contributed to this project: A. Narango, M. V. Parra and C. Dupe for help with the collection of the Colombian samples; N. Kalin and F. Laplace for discussions; A. Schiatti 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, D. Drury, C. Dupe, K. Gyaltsen, P. Jokinen, B. Lenz, S. Lewis, D. Lu, A. Narango, S. Ott, I. Padioleau, M. V. Parra, N. Patterson, A. Price, L. Suddendorf, S. Schorr, S. Sengamalay, J. Sullivan, F. Truelsen, T. Vickers, W. White and A. Schlattl for 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 of African ancestry in the Southeast United States; the Toscans 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 Ireland; the Palestinian populations in Spain; the Colombians in Medellin, Colombia; and the Puerto Ricans in Puerto Rico. This research was supported in part by Welcome Trust grants WT048851 to R.M.D.H. and C.T.S.; WT090532/2/09/2 to WT095525/2/09/2 and WT095525/1/11 to P.D.; WT086008/2/09 to WT090532/2/09/2 to G.A.M.; WT089250/2/09/2 to I.M.; WT085532/2/09/2 to P.F.; Medical Research Council grant G090074/7/I0/1070 to G.A.M.; British Heart Foundation grant RG09/01/2806 to C.A.A.; the National Basic Research Program of China (973 program no. 2011CB909201, 2011CB909202 and 2011CB909203); the BBSRC grant (2012AA020040); the National Natural Science Foundation of China (30890032, 31171337 and 30572089) to G.A.M.; the French National Research Agency (CNRS) to G.R.A. and T.Z.; the Swiss National Science Foundation 310003A_13043 to G.A.M.; the Dutch Research Centre for Genomics and Bioinformatics (C3X) to T.Z. and G.R.A.; the Shenzhen Key Laboratory of Transomics Biotechnologies (ZYSY201105170397A) to G.A.M.; the National Institutes of Health (NIH) grants U01HG5219, R2CHGS581 and R01HM84699 to G.R.A.; R01HG4719 and R01HG3698 to G.T.M.; RC2HG5552 and U01HG6513 to G.R.A. and P.J.H.; the Max Planck Society; Swiss National Science Foundation 310005133 to G.R.A.; the National Basic Research Program of China (973 program no. 2009CB906200) and 3137/3005/009A in the online version of the paper.
