A global reference for human genetic variation

Wang, Jun; Feng, Qiang; Fang, Xiaodong; Guo, Xiaosen; Jian, Min; Jiang, Hui; Liu, Xiao; Fang, Lin; Cao, Hongzhi; Jespersen, Jakob Berg

Published in:
Nature

DOI:
10.1038/nature15393

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):

Download date: 29. apr. 2019
A global reference for human genetic variation

The 1000 Genomes Project Consortium*

The 1000 Genomes Project set out to provide a comprehensive description of common human genetic variation by applying whole-genome sequencing to a diverse set of individuals from multiple populations. Here we report completion of the project, having reconstructed the genomes of 2,504 individuals from 26 populations using a combination of low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping. We characterized a broad spectrum of genetic variation, in total over 88 million variants (84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants), all phased onto high-quality haplotypes. This resource includes >99% of SNP variants with a frequency of >1% for a variety of ancestries. We describe the distribution of genetic variation across the global sample, and discuss the implications for common disease studies.
The total number of observed non-reference sites differs greatly among populations (Fig. 1b). Individuals from African ancestry populations harbour the greatest numbers of variant sites, as predicted by the out-of-Africa model of human origins. Individuals from recently admixed populations show great variability in the number of variants, roughly proportional to the degree of recent African ancestry in their genomes.

The majority of variants in the data set are rare: ~64 million autosomal variants have a frequency <0.5%, ~12 million have a frequency between 0.5% and 5%, and only ~8 million have a frequency >5% (Extended Data Fig. 3a). Nevertheless, the majority of variants observed in a single genome are common; just 40,000 to 200,000 of the variants in a typical genome (1–4%) have a frequency <0.5% (Fig. 1c and Extended Data Fig. 3b). As such, we estimate that improved rare variant discovery by deep sequencing our entire sample would at least double the total number of variants in our sample but increase the number of variants in a typical genome by only ~20,000 to 60,000.

Putatively functional variation

When we restricted analyses to the variants most likely to affect gene function, we found a typical genome contained 149–182 sites with protein truncating variants, 10,000 to 12,000 sites with peptide-sequence-altering variants, and 459,000 to 565,000 variant sites overlapping known regulatory regions (untranslated regions (UTRs), promoters, insulators, enhancers, and transcription factor binding sites). African genomes were consistently at the high end of these ranges. The number of alleles associated with a disease or phenotype in each genome did not follow this pattern of increased diversity in Africa (Extended Data Fig. 4): we observed ~2,000 variants per genome associated with complex traits through genome-wide association studies (GWAS) and 24–30 variants per genome implicated in rare disease through ClinVar; with European ancestry genomes at the high-end of these counts. The magnitude of this difference is unlikely to be explained by demography\textsuperscript{10,11}, but instead reflects the ethnic bias of current genetic studies. We expect that improved characterization of the clinical and phenotypic consequences of non-European alleles will enable better interpretation of genomes from all individuals and populations.

Sharing of genetic variants among populations

Systematic analysis of the patterns in which genetic variants are shared among individuals and populations provides detailed accounts of population history. Although most common variants are shared across the world, rarer variants are typically restricted to closely related populations (Fig. 1a); 86% of variants were restricted to a single continental group. Using a maximum likelihood approach\textsuperscript{12}, we estimated the proportion of each genome derived from several putative ‘ancestral populations’ (Fig. 2a and Extended Data Fig. 5).
This analysis separates continental groups, highlights their internal substructure, and reveals genetic similarities between related populations. For example, east–west clines are visible in Africa and East Asia, a north–south cline is visible in Europe, and European, African, and Native-American admixture is visible in genomes sampled in the Americas.

To characterize more recent patterns of shared ancestry, we first focused on variants observed on just two chromosomes (sample frequency of 0.04%), the rarest shared variants within our sample, and known as f2 variants. As expected, these variants are typically geographically restricted and much more likely to be shared between individuals in the same population or continental group, or between populations with known recent admixture (Extended Data Fig. 6a, b). Analysis of shared haplotype lengths around f2 variants suggests a median common ancestor ~296 generations ago (7,410 to 8,892 years ago; Extended Data Fig. 6c, d), although those confined within a population tend to be younger, with a shared common ancestor ~143 generations ago (3,570 to 4,284 years ago)13.

**Insights about demography**

Modelling the distribution of variation within and between genomes can provide insights about the history and demography of our ancestor populations14. We used the pairwise sequentially Markovian coalescent (PSMC)14 method to characterize the effective population size (\(N_e\)) of the ancestral populations (Fig. 2b and Extended Data Fig. 7). Our results show a shared demographic history for all humans beyond ~150,000 to 200,000 years ago. Further, they show that European, Asian and American populations shared strong and sustained bottlenecks, all with \(N_e < 1,500\), between 15,000 to 20,000 years ago. In contrast, the bottleneck experienced by African populations during the same time period appears less severe, with \(N_e > 4,250\). These bottlenecks were followed by extremely rapid inferred population growth in non-African populations, with notable exceptions including the PEL, MXL and FIN.

Due to the shared ancestry of all humans, only a modest number of variants show large frequency differences among populations. We observed 762,000 variants that are rare (defined as having frequency <0.5%) within the global sample but much more common (>5% frequency) in at least one population (Fig. 3a). Several populations have relatively large numbers of these variants, and these are typically genetically or geographically distinct within their continental group (LWK in Africa, PEL in the Americas, JPT in East Asia, FIN in Europe, and GIH in South Asia; see Supplementary Table 5). Drifited variants within such populations may reveal phenotypic associations that would be hard to identify in much larger global samples15.

Analysis of the small set of variants with large frequency differences between closely related populations can identify targets of recent, localized adaptation. We used the FST-based population branch statistic (PBS)16 to identify genes with strong differentiation between pairs of populations in the same continental group (Fig. 3b). This approach reveals a number of previously identified selection signals (such as SLC24A5 associated with skin pigmentation17, HERC2 associated with eye colour18, LCT associated with lactose tolerance, and the FADS cluster that may be associated with dietary fat sources19). Several potentially novel selection signals are also highlighted (such as TRBV9, which appears particularly differentiated in South Asia, PRICKLE4, differentiated in African and South Asian populations, and a number of genes in the immunoglobulin cluster, differentiated in East Asian populations; Extended Data Fig. 8), although at least some of these signals may result from somatic rearrangements (for example, via V(D)J recombination) and differences in cell type composition among the sequenced samples. Nonetheless, the relatively small number of genes showing strong differentiation between closely related populations highlights the rarity of strong selective sweeps in recent human evolution20.

**Sharing of haplotypes and imputation**

The sharing of haplotypes among individuals is widely used for imputation in GWAS, a primary use of 1000 Genomes data. To assess imputation based on the phase 3 data set, we used Complete Genomics data for 9 or 10 individuals from each of 6 populations (CEU, CHS, LWK, PEL, PJL, and YRI). After excluding these individuals from the reference panel, we imputed genotypes across the genome using sites on a typical one million SNP microarray. The squared correlation between imputed and experimental genotypes was >95% for common variants in each population, decreasing gradually with minor allele frequency (Fig. 4a). Compared to phase 1, rare variation imputation improved considerably, particularly for newly sampled populations (for example, PEL and PJL, Extended Data Fig. 9a). Improvements in imputations restricted to overlapping samples suggest approximately equal contributions from greater genotype and sequence quality and from increased sample size (Fig. 4a, inset). Imputation accuracy is now similar for bi-allelic SNPs, bi-allelic indels, multi-allelic SNPs, and sites where indels and SNPs overlap, but slightly reduced for multi-allelic indels, which typically map to regions of low-complexity sequence and are much harder to genotype and phase (Extended Data Fig. 9b). Although imputation of rare variation remains challenging, it appears to be
Table 1 | Median autosomal variant sites per genome

<table>
<thead>
<tr>
<th></th>
<th>AFR</th>
<th>AMR</th>
<th>EAS</th>
<th>EUR</th>
<th>SAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samples</td>
<td>661</td>
<td>347</td>
<td>504</td>
<td>503</td>
<td>489</td>
</tr>
<tr>
<td>Mean coverage</td>
<td>8.2</td>
<td>7.6</td>
<td>7.7</td>
<td>7.4</td>
<td>8.0</td>
</tr>
<tr>
<td>SNPs</td>
<td>4.31M</td>
<td>14.5k</td>
<td>3.64M</td>
<td>12.0k</td>
<td>3.55M</td>
</tr>
<tr>
<td>Indels</td>
<td>625k</td>
<td>557k</td>
<td>-</td>
<td>940k</td>
<td>7</td>
</tr>
<tr>
<td>Large deletions</td>
<td>1.1k</td>
<td>5</td>
<td>949</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>CNVs</td>
<td>170</td>
<td>1</td>
<td>153</td>
<td>1</td>
<td>182</td>
</tr>
<tr>
<td>MEI (Alu)</td>
<td>1.03k</td>
<td>0</td>
<td>845</td>
<td>0</td>
<td>899</td>
</tr>
<tr>
<td>MEI (L1)</td>
<td>138</td>
<td>0</td>
<td>118</td>
<td>0</td>
<td>130</td>
</tr>
<tr>
<td>MEI (SVA)</td>
<td>52</td>
<td>0</td>
<td>44</td>
<td>0</td>
<td>56</td>
</tr>
<tr>
<td>MEI (MT)</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Inversions</td>
<td>12</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Nonsense</td>
<td>12.2k</td>
<td>139</td>
<td>10.4k</td>
<td>121</td>
<td>10.2k</td>
</tr>
<tr>
<td>Synon</td>
<td>13.8k</td>
<td>78</td>
<td>11.4k</td>
<td>67</td>
<td>11.2k</td>
</tr>
<tr>
<td>Intron</td>
<td>2.06M</td>
<td>7.33k</td>
<td>1.72M</td>
<td>6.12k</td>
<td>1.68M</td>
</tr>
<tr>
<td>UTR</td>
<td>37.2k</td>
<td>168</td>
<td>30.8k</td>
<td>136</td>
<td>30.0k</td>
</tr>
<tr>
<td>Promoter</td>
<td>102k</td>
<td>430</td>
<td>84.3k</td>
<td>332</td>
<td>81.6k</td>
</tr>
<tr>
<td>Insulator</td>
<td>70.9k</td>
<td>248</td>
<td>59.0k</td>
<td>199</td>
<td>57.7k</td>
</tr>
<tr>
<td>Enhancer</td>
<td>354k</td>
<td>1.32k</td>
<td>295k</td>
<td>1.05k</td>
<td>289k</td>
</tr>
<tr>
<td>TFBSs</td>
<td>927</td>
<td>4</td>
<td>759</td>
<td>3</td>
<td>748</td>
</tr>
<tr>
<td>Filtered LoF</td>
<td>182</td>
<td>4</td>
<td>152</td>
<td>3</td>
<td>153</td>
</tr>
<tr>
<td>HGMD-DM</td>
<td>20</td>
<td>0</td>
<td>18</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>GWAS</td>
<td>2.00k</td>
<td>0</td>
<td>2.07k</td>
<td>0</td>
<td>1.99k</td>
</tr>
<tr>
<td>ClinVar</td>
<td>28</td>
<td>0</td>
<td>30</td>
<td>1</td>
<td>24</td>
</tr>
</tbody>
</table>

See Supplementary Table 1 for continental population groupings. CNVs, copy-number variants; HGMD-DM, Human Gene Mutation Database disease mutations; k, thousand; LoF, loss-of-function; M, million; MEI, mobile element insertions.

The performance of imputation and GWAS studies depends on the local distribution of linkage disequilibrium (LD) between nearby variants. Controlling for sample size, the decay of LD as a function of physical distance is fastest in African populations and slowest in East Asian populations (Extended Data Fig. 10). To evaluate how these differences influence the resolution of genetic association studies and,Mouse. Rat. Drosophila. Arabidopsis. 

Figure 2 | Population structure and demography. a. Population structure inferred using a maximum likelihood approach with 8 clusters. b. Changes to effective population sizes over time, inferred using PSMC. Lines represent the within-population median PSMC estimate, smoothed by fitting a cubic spline passing through bin midpoints.

©2015 Macmillan Publishers Limited. All rights reserved
Figure 3 | Population differentiation. a, Variants found to be rare (<0.5%) within the global sample, but common (>5%) within a population. b, Genes showing strong differentiation between pairs of closely related populations.

In particular, their ability to identify a narrow set of candidate functional variants, we evaluated the number of tagging variants ($r^2 > 0.8$) for a typical variant in each population. We find that each common variant typically has over 15–20 tagging variants in non-African populations, but only about 8 in African populations (Fig. 4b). At lower frequencies, we find 3–6 tagging variants with 100 kb of variants with frequency <0.5%, and differences in the number of tagging variants between continental groups are less marked.

Among variants in the GWAS catalogue (which have an average frequency of 26.6% in project haplotypes), the number of proxies averages 14.4 in African populations and 30.3–44.4 in other continental groupings (Supplementary Table 10). The potential value of eQTL variants that are SNPs and indels, as discovered in 69 samples from the population in which the maximum value was achieved.

Figure 4 | Imputation and eQTL discovery. a, Imputation accuracy as a function of allele frequency for six populations. The insert compares imputation accuracy between phase 3 and phase 1, using all samples (solid lines) and intersecting samples (dashed lines). b, The average number of tagging variants ($r^2 > 0.8$) as a function of physical distance for common (top), low frequency (middle), and rare (bottom) variants. c, The proportion of top eQTL variants that are SNPs and indels, as discovered in 69 samples from each population. d, The percentage of eQTLs in TFBS, having performed discovery in the first population, and fine mapped by including an additional 69 samples from a second population ($*P < 0.01$, **$P < 0.001$, ***$P < 0.0001$, McNemar’s test). The diagonal represents the percentage of eQTLs in TFBS using the original discovery sample.
multi-population fine-mapping is illustrated by the observation that the number of proxies shared across all populations is only 8.2 and, furthermore, that 34.9% of GWAS catalogue variants have no proxy shared across all continental groupings.

To further assess prospects for fine-mapping genetic association signals, we performed expression quantitative trait loci (eQTL) discovery at 17,667 genes in 69 samples from each of 6 populations (CEU, CHB, GH, JPT, LWK, and YRI)\(^2\). We identified eQTLs for 3.285 genes at 5% FDR (average 1,265 genes per population). Overall, a typical eQTL signal comprised 67 associated variants, including an indel as one of the top associated variants 26–40% of the time (Fig. 4c). Within each discovery population, 17.5–19.5% of top eQTL variants overlapped annotated transcription factor binding sites (TFBSs), consistent with the idea that a substantial fraction of eQTL polymorphisms are TFBS polymorphisms. Using a meta-analysis approach to combine pairs of populations, the proportion of top eQTL variants overlapping TFBSs increased to 19.2–21.6% (Fig. 4d), consistent with improved localization. Including an African population provided the greatest reduction in the count of associated variants and the greatest increase in overlap between top variants and TFBSs.

**Discussion**

Over the course of the 1000 Genomes Project there have been substantial advances in sequence data generation, archiving and analysis. Primary sequence data production improved with increased read length and depth, reduced per-base errors, and the introduction of paired-end sequencing. Sequence analysis methods improved with the development of strategies for identifying and filtering poor-quality data, for more accurate mapping of sequence reads (particularly in repetitive regions), for exchanging data between analysis tools and enabling ensemble analyses, and for capturing more diverse types of variants. Importantly, each release has examined larger numbers of individuals, aiding population-based analyses that identify and leverage shared haplotypes during genotyping. Whereas our first analyses produced high-confidence short-variant calls for 80–85% of the reference genome\(^1\), our newest analyses reach ~96% of the genome using the same metrics, although our ability to accurately capture structural variation remains more limited\(^3\). In addition, the evolution of sequencing, analysis and filtering strategies means that our results are not a simple superset of previous analysis. Although the number of characterized variants has more than doubled relative to phase 1, ~2.3 million previously described variants are not included in the current analysis; most missing variants were rare or marked as low quality: 1.6 million had frequency <0.5% and may be missing from our current read set, while the remainder were removed by our filtering processes.

These same technical advances are enabling the application of whole genome sequencing to a variety of medically important samples. Some of these studies already exceed the 1000 Genomes Project in size\(^4\–\(^8\), but the results described here remain a prime resource for studies of genetic variation for several reasons. First, the 1000 Genomes Project samples provide a broad representation of human genetic variation—in contrast to the bulk of complex disease studies in humans, which primarily study European ancestry samples and which, as we show, fail to capture functionally important variation in other populations. Second, the project analyses incorporate multiple analysis strategies, callsets and variant types. Although such ensemble analyses are cumbersome, they provide a benchmark for what can be achieved and a yardstick against which more practical analysis strategies can be evaluated. Third, project samples and data resulting from them can be shared broadly, enabling sequencing strategies and analysis methods to be compared easily on a benchmark set of samples. Because of the wide availability of the data and samples, these samples have been and will continue to be used for studying many molecular phenotypes. Thus, we predict that the samples will accumulate many types of data that will allow connections to be drawn between variants and both molecular and disease phenotypes.

**Online Content**

Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 12 May; accepted 20 August 2015.**


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

**Acknowledgements** We thank the many people who were generous with contributing their samples to the project: the African Caribbean in Barbados; Bengali in Bangladesh; British in England and Scotland; Chinese Dai in Xishuangbanna, China; Colombians in Medellin, Colombia; Ewan in Nigeria; Finnish in Finland; Gambian in Western Division – Mandinka; Gujarati Indians in Houston, Texas, USA; Han Chinese in Beijing, China; Iberian populations in Spain; Indian Telugu in the UK; Japanese in Tokyo, Japan; Kinh in Ho Chi Minh City, Vietnam; Luhya inWEBuye, Kenya; Mend in Sierra Leone; people with African ancestry in the southwest USA; people with Mexican ancestry in Los Angeles, California, USA; Peruvians in Lima, Peru; Puerto Ricans in Puerto Rico; Punjabi in Lahore, Pakistan; southern Han Chinese; Sri Lankan Tamil in the UK; Toscans in Italy; Utah residents (CEPH) with northern and western European ancestry; and Yoruba in Ibadan, Nigeria. Many thanks to the people who contributed to this project: P. Maul, T. Maul, and C. Foster; Z. Chong, X. Fan, W. Zhou, and T. Chen; N. Sengamalay, S. Ott, L. Sadzwicz, J. Liu, and L. Tailon; L. Merson; O. Folarin; D. Asogun, O. Ikpwonmosa, E. Philomena, G. Akpede, S. Okhobgenin, and O. Omoniwa; the staff of the Institute of Lassa Fever Research and Control (ILFRC), Irrua Specialist Teaching Hospital, Irrua, Edo State, Nigeria; A. Schlattl and T. Zichner; S. Lewis, E. Appelbaum, and L. Fulton; A. Yurovsky and I. Padioleau; N. Kaelin and F. Laplace; E.E.E. is an investigator of the Howard Hughes Medical Institute.

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

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.A. (adam.auton@gmail.com) or G.R.A. (goncalo@umich.edu).

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported licence. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons licence, users will need to obtain permission from the licence holder to reproduce the material. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-sa/3.0/
Extended Data Figure 1 | Summary of the callset generation pipeline. Boxes indicate steps in the process and numbers indicate the corresponding section(s) within the Supplementary Information.
Extended Data Figure 2 | Power of discovery and heterozygote genotype discordance. a, The power of discovery within the main data set for SNPs and indels identified within an overlapping sample of 284 genomes sequenced to high coverage by Complete Genomics (CG), and against a panel of >60,000 haplotypes constructed by the Haplotype Reference Consortium (HRC). To provide a measure of uncertainty, one curve is plotted for each chromosome. b, Improved power of discovery in phase 3 compared to phase 1, as assessed in a sample of 170 Complete Genomics genomes that are included in both phase 1 and phase 3. c, Heterozygote discordance in phase 3 for SNPs, indels, and SVs compared to 284 Complete Genomics genomes. d, Heterozygote discordance for phase 3 compared to phase 1 within the intersecting sample. e, Sensitivity to detect Complete Genomics SNPs as a function of sequencing depth. f, Heterozygote genotype discordance as a function of sequencing depth, as compared to Complete Genomics data.
Extended Data Figure 3 | Variant counts.  
a. The number of variants within the phase 3 sample as a function of alternative allele frequency.  
b. The average number of detected variants per genome with whole-sample allele frequencies <0.5% (grey bars), with the average number of singletons indicated by colours.
Extended Data Figure 4 | The standardized number of variant sites per genome, partitioned by population and variant category. For each category, z-scores were calculated by subtracting the mean number of sites per genome (calculated across the whole sample), and dividing by the standard deviation.

From left: sites with a derived allele, synonymous sites with a derived allele, nonsynonymous sites with a derived allele, sites with a loss-of-function allele, sites with a HGMD disease mutation allele, sites with a ClinVar pathogenic variant, and sites carrying a GWAS risk allele.
Extended Data Figure 5 | Population structure as inferred using the admixture program for $K = 5$ to 12.
Extended Data Figure 6 | Allelic sharing. a, Genotype covariance (above diagonal) and sharing of $f_2$ variants (below diagonal) between pairs of individuals. b, Quantification of average $f_2$ sharing between populations. Each row represents the distribution of $f_2$ variants shared between individuals from the population indicated on the left to individuals from each of the sampled populations. c, The average number of $f_2$ variants per haploid genome. d, The inferred age of $f_2$ variants, as estimated from shared haplotype lengths, with black dots indicating the median value.
Extended Data Figure 7 | Unsmoothed PSMC curves. a, The median PSMC curve for each population. b, PSMC curves estimated separately for all individuals within the 1000 Genomes sample. c, Unsmoothed PSMC curves comparing estimates from the low coverage data (dashed lines) to those obtained from high coverage PCR-free data (solid lines). Notable differences are confined to very recent time intervals, where the additional rare variants identified by deep sequencing suggest larger population sizes.
Extended Data Figure 8 | Genes showing very strong patterns of differentiation between pairs of closely related populations within each continental group. Within each continental group, the maximum PBS statistic was selected from all pairwise population comparisons within the continental group against all possible out-of-continent populations. Note the x axis shows the number of polymorphic sites within the maximal comparison.
Extended Data Figure 9 | Performance of imputation. a, Performance of imputation in 6 populations using a subset of phase 3 as a reference panel (n = 2,445), phase 1 (n = 1,065), and the corresponding data within intersecting samples from both phases (n = 1,006). b, Performance of imputation from phase 3 by variant class.
Extended Data Figure 10 | Decay of linkage disequilibrium as a function of physical distance. Linkage disequilibrium was calculated around 10,000 randomly selected polymorphic sites in each population, having first thinned each population down to the same sample size (61 individuals). The plotted line represents a 5 kb moving average.