A global reference for human genetic variation

The 1000 Genomes Project Consortium[1]

[1]A full list of authors can be found at the end of the document

Last Modified: 7/28/15 10:23 AM

Summary

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, totaling over 88 million variants (84.7 million SNPs, 3.6 million short indels, and 60 thousand 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.

Main Text

The 1000 Genomes Project has already elucidated the properties and distribution of common and rare variation, provided insights into the processes that shape genetic diversity, and advanced understanding of disease biology[1,2]. This resource provides a benchmark for surveys of human genetic variation and constitutes a key component for human genetic studies, by enabling array design[3,4], genotype imputation[5], cataloguing of variants in regions of interest, and filtering of likely neutral variants[6,7].

In this final phase, individuals were sampled from 26 populations in Africa (AFR), East Asia (EAS), Europe (EUR), South Asia (SAS), and the Americas (AMR) (Figure 1A; see Supplementary Table 1 for population descriptions and abbreviations). All individuals were sequenced using both whole-genome sequencing (mean depth = 7.4X) and targeted exome sequencing (mean depth = 65.7X). In addition, individuals and available first-degree relatives (generally adult offspring) were genotyped using high-density SNP microarrays. This provided a cost-effective means to discover genetic variants and estimate individual genotypes and haplotypes[1,2].

Dataset Overview

In contrast to earlier phases of the project, we expanded analysis beyond bi-allelic events to include multi-allelic SNPs, indels, and a diverse set of structural variants (SV). Variant discovery used an ensemble of 24 sequence analysis tools (Supplementary Table 2), and machine-learning classifiers to separate high-quality variants from potential false positives, balancing sensitivity and specificity. Construction of haplotypes started with estimation of long-range phased haplotypes using array genotypes for project participants and, where available, their first degree relatives; continued with the addition of bi-allelic SNPs and indels that were analyzed jointly to improve these haplotypes; and concluded with the placement of multi-allelic and structural variants onto the haplotype scaffold one at a time (Box 1). Overall, we discovered, genotyped, and phased 88 million variant sites (Supplementary Table 3). The project has now contributed or validated 80 of the 100 million variants in the public dbSNP catalog (version 141 includes 40 million SNPs and indels newly contributed by this analysis). These novel variants especially enhance our catalog of genetic variation within South Asian (which account for 24% of novel variants) and African populations (28% of novel variants).

To control the false discovery rate (FDR) of SNPs and indels at <5%, a variant quality score threshold was defined using high depth (>30X) PCR-free sequence data generated for one individual per population. For structural variants, additional orthogonal methods were used for confirmation, including microarrays and long read sequencing, resulting in FDR <5% for deletions, duplications, multi-allelic copy-number variants, Alu and L1 insertions, and <20% for inversions, SVA (SINE/*VNTR/Alu*) composite retrotransposon insertions and NUMTs[8] (nuclear

mitochondrial DNA variants). To evaluate variant discovery power and genotyping accuracy, we also generated deep Complete Genomics data (mean depth = 47X) for 427 individuals (129 mother-father-child trios, 12 parent-child duos, and 16 unrelateds). We estimate the power to detect SNPs and indels to be >95% and >80%, respectively, for variants with sample frequency of at least 0.5%, rising to >99% and >85% for frequencies >1% (Supplementary Figures 1 and 2). At lower frequencies, comparison with >60,000 European haplotypes from the Haplotype Reference Consortium[9] suggests 75% power to detect SNPs with frequency of 0.1%. Furthermore, we estimate heterozygous genotype accuracy at 99.4% for SNPs and 99.0% for indels (Supplementary Table 4), a three-fold reduction in error rates compared to our previous release[2], resulting from the larger sample size, improvements in sequence data accuracy, and genotype calling and phasing algorithms.

A Typical Genome

We find that a typical genome differs from the reference human genome at 4.09 - 5.02 million sites (Figure 1B; Table 1). While >99.9% of variants consist of SNPs and short indels, structural variants affect more bases: the typical genome contains an estimated 2,100-2,500 structural variants (~1,000 large deletions, ~160 copy-number variants, ~915 Alu insertions, ~128 L1 insertions, ~51 SVA insertions, ~4 NUMTs, and ~10 inversions), affecting ~20 million bases of sequence.

The total number of observed non-reference sites differs greatly among populations (Figure 1B). Individuals from African ancestry populations harbor 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 dataset are rare: ~64 million autosomal variants have frequency <0.5%, ~12 million have frequency between 0.5% and 5%, and only ~8 million have frequency >5% (Supplementary Figure 3A). Nevertheless, the majority of variants observed in a single genome are common: just 40 – 200 thousand of the variants in a typical genome (1-4%) have frequency <0.5% (Figure 1C, Supplementary Figure 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 – 60 thousand.

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 – 12 thousand sites with peptide sequence altering variants, and 459 – 565 thousand variant sites overlapping known regulatory 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 (Supplementary Figure 4): we observed ~2,000 variants per genome associated with complex traits through 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[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 globe, rarer variants are typically restricted to closely related populations (Figure 1A); 86% of variants were restricted to a single continental group. Using a maximum likelihood approach[12], we estimated the proportion of each genome derived from several putative 'ancestral populations' (Figure 2A; Supplementary Figure 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 f_2 variants[2]. 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 (Supplementary Figures 6A and B). Analysis of shared haplotype lengths around f_2 variants suggests a median common ancestor ~296 generations ago (7,410 – 8,892 years ago; Supplementary Figures 6C and D), although those confined within a population tend to be younger, with a shared common ancestor ~143 generations ago (3,570 – 4,284 years ago)[13].

Insights About Demography

Modeling the distribution of variation within and between genomes can provide insights about the history and demography of our ancestor populations[14]. We used the Pairwise Sequentially Markovian Coalescent (PSMC)[14] method to characterize the effective population size (N_e) of the ancestral populations (Figure 2B; Supplementary Figures 7 and 8). Our results clearly show a shared demographic history for all humans beyond ~150 – 200 thousand years ago. Further, they show

that European, Asian and American populations shared strong and sustained bottlenecks, all with $N_e < 1500$, between 15 - 20 thousand 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 (Figure 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 America, JPT in East Asia, FIN in Europe, and GIH in South Asia; see Supplementary Table 5). Drifted variants within such populations may reveal phenotypic associations that would be hard to identify in much larger global samples[15].

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 F_{ST}–based population branch statistic (PBS)[16] to identify genes with strong differentiation between pairs of populations in the same continental group (Figure 3B). This approach reveals a number of previously identified selection signals (such as SLC24A5 associated with skin pigmentation[17], HERC2 associated with eye color[18], LCT associated with lactose tolerance, and the FADS cluster that may be associated with dietary fat sources[19]). 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; Supplementary Figure 9), although at least some of these signals may result from somatic rearrangements (e.g. 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 evolution[20].

Sharing of Haplotypes and Imputation

The sharing of haplotypes among individuals is widely used for imputation in genome-wide association studies (GWAS), a primary use of 1000 Genomes data. To assess imputation based on the Phase 3 dataset, we used Complete Genomics data for 9-10 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 1M SNP microarray. The squared correlation between imputed and experimental genotypes was >95% for common variants in each population, decreasing gradually with minor allele frequency (Figure 4A). Compared to Phase 1, rare variation imputation improved considerably,

particularly for newly sampled populations (e.g. PEL and PJL, Supplementary Figure 10A). Improvements in imputations restricted to overlapping samples suggest approximately equal contributions from greater genotype and sequence quality and from increased sample size (Figure 4A insert). 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 (Supplementary Figure 10B). While imputation of rare variation remains challenging, it appears to be most accurate in African ancestry populations, where greater genetic diversity results in a larger number of haplotypes and improves the chances of rare variants being tagged by a characteristic haplotype.

Resolution of Genetic Association Studies

To evaluate the impact of our new reference panel on GWAS, we re-analyzed a previous study of age-related macular degeneration (AMD) totaling 2,157 cases and 1,150 controls[21]. We imputed 17.0M genetic variants with estimated $R^2 > 0.3$, compared to 14.1M variants using Phase 1, and only 2.4M SNPs using HapMap2. Compared to Phase 1, the number of imputed common and intermediate frequency variants increased by 7%, while the number of rare variants increased by >50%, and the number of indels increased by 70% (Supplementary Table 6). We permuted case-control labels to estimate a genome-wide significance threshold of $p < \sim 1.5 \times 10^{-8}$, which corresponds to ~3 million independent variants and is more stringent than the traditional threshold of 5×10^{-8} (Supplementary Table 7). In practice, significance thresholds must balance the false positives and false negatives[22-24]. We recommend thresholds for strict control of false positives be determined using permutations, and expect these to become more stringent thresholds as larger sample sizes, more diverse populations and/or direct sequencing are used. After imputation, five independent signals in four previously reported AMD loci[25-28] reached genome-wide significance (Supplementary Table 8). When we examined each of these to define a set of potentially causal variants using a Bayesian Credible set approach[29], lists of potentially functional variants were ~4x larger than in HapMap2-based analysis and 7% larger than in analyses based on Phase 1 (Supplementary Table 9). In one locus, the most strongly associated variant was now a structural variant (estimated imputation R^2 0.89) that previously could not be imputed, consistent with some functional studies[30]. Deep catalogues of potentially functional variants will help ensure that downstream functional analyses include the true candidate variants, and will aid analyses that integrate complex disease associations with functional genomic elements[31].

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 (Supplementary Figure 11). To evaluate how these differences influence the resolution of genetic association studies and, 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 (Figure 4B). While only ~3-6 tagging variants are typically found within 100 kb of variants with frequency <0.5%, the differences between the continental groups are considerably less marked.

Among variants in the GWAS catalog (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 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 catalog 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, GIH, JPT, LWK, and YRI)[32]. 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 26-40% of the time (Figure 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% (Figure 4D), consistent with improved localization. Including an African population provided the greatest reduction in the count of associated variants and increased overlap between top variants and TFBSs.

Discussion

The course of the 1000 Genomes Project has witnessed 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[33]. In addition, the evolution of sequencing, analysis and filtering strategies means that our results are not a simple superset of previous analysis.

While 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 of which 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[34-36], 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. While 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 the samples will accumulate many types of data that will allow connections to be drawn between variants and both molecular and disease phenotypes.

Box 1: Building a haplotype scaffold

(Box Figure 1)

To construct high quality haplotypes that integrate multiple variant types, we adopted a staged approach[37]. 1) A high-quality 'haplotype scaffold' was constructed using statistical methods applied to SNP microarray genotypes and, where available, genotypes for first degree relatives (available for ~52% of samples; Supplementary Table 11)[38]. 2a) Variant sites were identified using a combination of bioinformatic tools and pipelines to define a set of high-confidence bi-allelic variants, including both SNPs and indels (white triangles), which were jointly imputed onto the haplotype scaffold. 2b) Multi-allelic SNPs, indels, and complex variants (represented by yellow shapes, or variation in copy number) were placed onto the haplotype scaffold one-at-a-time, exploiting the local linkage disequilibrium information but leaving haplotypes for other variants undisturbed[39]. 3) The biallelic and multi-allelic haplotypes were merged into a single haplotype representation. This multi-stage approach allows the long-range structure of the haplotype scaffold to be maintained while including more complex types of variation. Comparison to haplotypes constructed from fosmids suggests the average distance between phasing errors is ~1062 kb, while typical phasing errors are ~37 kb (Supplementary Table 12).

Methods Summary

Details concerning sample collection, data generation, data processing, and analysis are in the Supplementary Information. Supplementary Figure 12 summarizes the process and points to the relevant supplementary sections.

References

1 The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).

2 The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).

3 Voight, B. F. *et al.* The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS genetics* **8**, e1002793, doi:10.1371/journal.pgen.1002793 (2012).

4 Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nature genetics* **43**, 1193-1201, doi:10.1038/ng.998 (2011).

5 Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature genetics* **44**, 955-959, doi:10.1038/ng.2354 (2012).

6 Xue, Y. *et al.* Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *American journal of human genetics* **91**, 1022-1032, doi:10.1016/j.ajhg.2012.10.015 (2012).

7 Jung, H., Bleazard, T., Lee, J. & Hong, D. Systematic investigation of cancer-associated somatic point mutations in SNP databases. *Nature biotechnology* **31**, 787-789, doi:10.1038/nbt.2681 (2013).

8 Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Submitted to Nature* (2015).

9 The Haplotype Reference Consortium. <http://www.haplotype-reference-consortium.org/> (2015).

10 Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nature genetics* **46**, 220-224, doi:10.1038/ng.2896 (2014).

11 Do, R. *et al.* No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nature genetics* **47**, 126-131, doi:10.1038/ng.3186 (2015).

12 Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research* **19**, 1655-1664, doi:10.1101/gr.094052.109 (2009).

13 Mathieson, I. & McVean, G. Demography and the age of rare variants. *PLoS genetics* **10**, e1004528, doi:10.1371/journal.pgen.1004528 (2014).

14 Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493-496, doi:10.1038/nature10231 (2011).
15 Moltke, I. *et al.* A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. *Nature* **512**, 190-193, doi:10.1038/nature13425 (2014).
16 Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75-78, doi:10.1126/science.1190371 (2010).
17 Lamason, R. L. *et al.* SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* **310**, 1782-1786, doi:10.1126/science.1116238 (2005).
18 Eiberg, H. *et al.* Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Human genetics* **123**, 177-187, doi:10.1007/s00439-007-0460-x (2008).
19 Mathias, R. A. *et al.* Adaptive evolution of the FADS gene cluster within Africa. *PloS one* **7**, e44926, doi:10.1371/journal.pone.0044926 (2012).
20 Hernandez, R. D. *et al.* Classic selective sweeps were rare in recent human evolution. *Science* **331**, 920-924, doi:10.1126/science.1198878 (2011).
21 Chen, W. *et al.* Genetic variants near TIMP3 and high-density lipoprotein-associated loci influence susceptibility to age-related macular degeneration. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 7401-7406, doi:10.1073/pnas.0912702107 (2010).
22 Wakefield, J. Bayes factors for genome-wide association studies: comparison with P-values. *Genet Epidemiol* **33**, 79-86, doi:10.1002/gepi.20359 (2009).
23 Wakefield, J. Commentary: Genome-wide significance thresholds via Bayes factors. *Int J Epidemiol* **41**, 286-291, doi:10.1093/ije/dyr241 (2012).
24 Sham, P. C. & Purcell, S. M. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet* **15**, 335-346, doi:10.1038/nrg3706 (2014).
25 Gold, B. *et al.* Variation in factor B (BF) and complement component 2 (C2) genes is associated with age-related macular degeneration. *Nature genetics* **38**, 458-462, doi:10.1038/ng1750 (2006).
26 Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385-389, doi:10.1126/science.1109557 (2005).
27 Rivera, A. *et al.* Hypothetical LOC387715 is a second major susceptibility gene for age-related macular degeneration, contributing independently of complement factor H to disease risk. *Hum Mol Genet* **14**, 3227-3236, doi:10.1093/hmg/ddi353 (2005).
28 Yates, J. R. *et al.* Complement C3 variant and the risk of age-related macular degeneration. *N Engl J Med* **357**, 553-561, doi:10.1056/NEJMoa072618 (2007).
29 Maller, J. B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature genetics* **44**, 1294-1301, doi:10.1038/ng.2435 (2012).

30 Fritsche, L. G. *et al.* Age-related macular degeneration is associated with an unstable ARMS2 (LOC387715) mRNA. *Nat. Genet.* **40**, 892-896, doi:ng.170 [pii]
10.1038/ng.170 (2008).
31 ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).
32 Stranger, B. E. *et al.* Patterns of cis regulatory variation in diverse human populations. *PLoS genetics* **8**, e1002639, doi:10.1371/journal.pgen.1002639 (2012).
33 Chaisson, M. J. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608-611, doi:10.1038/nature13907 (2015).
34 Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nature genetics* **47**, 435-444, doi:10.1038/ng.3247 (2015).
35 The UK10K Consortium. The UK10K project: rare variants in health and disease. *Nature* **To appear** (2015).
36 Sidore, C. *et al.* Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nature genetics* **To appear** (2015).
37 Delaneau, O., Marchini, J. & The 1000 Genomes Project Consortium. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat Commun* **5**, doi:10.1038/Ncomms4934 (2014).
38 O'Connell, J. *et al.* A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS genetics* **10**, e1004234, doi:10.1371/journal.pgen.1004234 (2014).
39 Menelaou, A. & Marchini, J. Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. *Bioinformatics* **29**, 84-91, doi:10.1093/bioinformatics/bts632 (2013).

Tables

	AFR		AMR		EAS		EUR		SAS	
Samples	661		347		504		503		489	
Mean Coverage	8.2		7.6		7.7		7.4		8.0	
	Var. Sites	**Singletons**	**Var. Sites**	**Singletons**	**Var. Sites**	**Singletons**	**Var. Sites**	**Singletons**	**Var. Sites**	**Singletons**
SNPs	4.31M	14.5k	3.64M	12.0k	3.55M	14.8k	3.53M	11.4k	3.60M	14.4k
Indels	625k	-	557k	-	546k	-	546k	-	556k	-
Large Deletions	1.1k	5	949	5	940	7	939	5	947	5
CNVs	170	1	153	1	158	1	157	1	165	1
MEI (Alu)	1.03k	0	845	0	899	1	919	0	889	0
MEI (LINE1)	138	0	118	0	130	0	123	0	123	0
MEI (SVA)	52	0	44	0	56	0	53	0	44	0
MEI (MT)	5	0	5	0	4	0	4	0	4	0
Inversions	12	0	9	0	10	0	9	0	11	0
NonSynon	12.2k	139	10.4k	121	10.2k	144	10.2k	116	10.3k	144
Synon	13.8k	78	11.4k	67	11.2k	79	11.2k	59	11.4k	78
Intron	2.06M	7.33k	1.72M	6.12k	1.68M	7.39k	1.68M	5.68k	1.72M	7.20k
UTR	37.2k	168	30.8k	136	30.0k	169	30.0k	129	30.7k	168
Promoter	102k	430	84.3k	332	81.6k	425	82.2k	336	84.0k	430
Insulator	70.9k	248	59.0k	199	57.7k	252	57.7k	189	59.1k	243
Enhancer	354k	1.32k	295k	1.05k	289k	1.34k	288k	1.02k	295k	1.31k
TFBS	927	4	759	3	748	4	749	3	765	3
Filtered LoF	182	4	152	3	153	4	149	3	151	3
HGMD-DM	20	0	18	0	16	1	18	2	16	0
GWAS	2.00k	0	2.07k	0	1.99k	0	2.08k	0	2.06k	0
ClinVar	28	0	30	1	24	0	29	1	27	1

Table 1: Median autosomal variant sites per genome. See Supplementary Table 1 for continental population groupings.

Main Figures

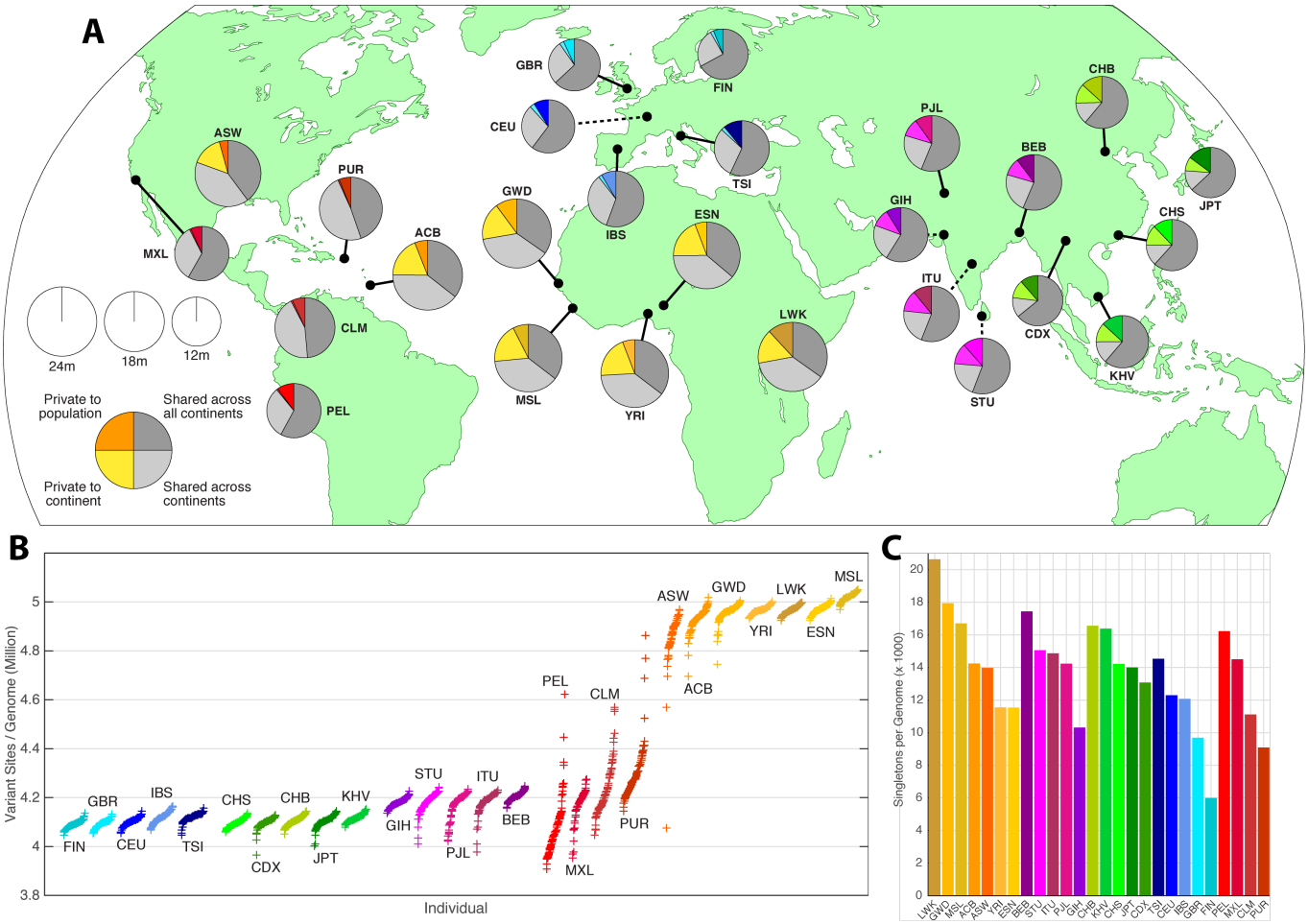


Figure 1: Population sampling. **A)** Polymorphic variants within sampled populations. The area of each pie is proportional to the number of polymorphisms within a population. Pies are divided into four slices, representing variants private to a population (darker color unique to population), private to a continental area (lighter color shared across continental group), shared across continental areas (light grey), and shared across all continents (dark grey). Dashed lines indicate populations sampled outside of their ancestral continental region. **B)** The number of variant sites per genome. **C)** The average number of singletons per genome.

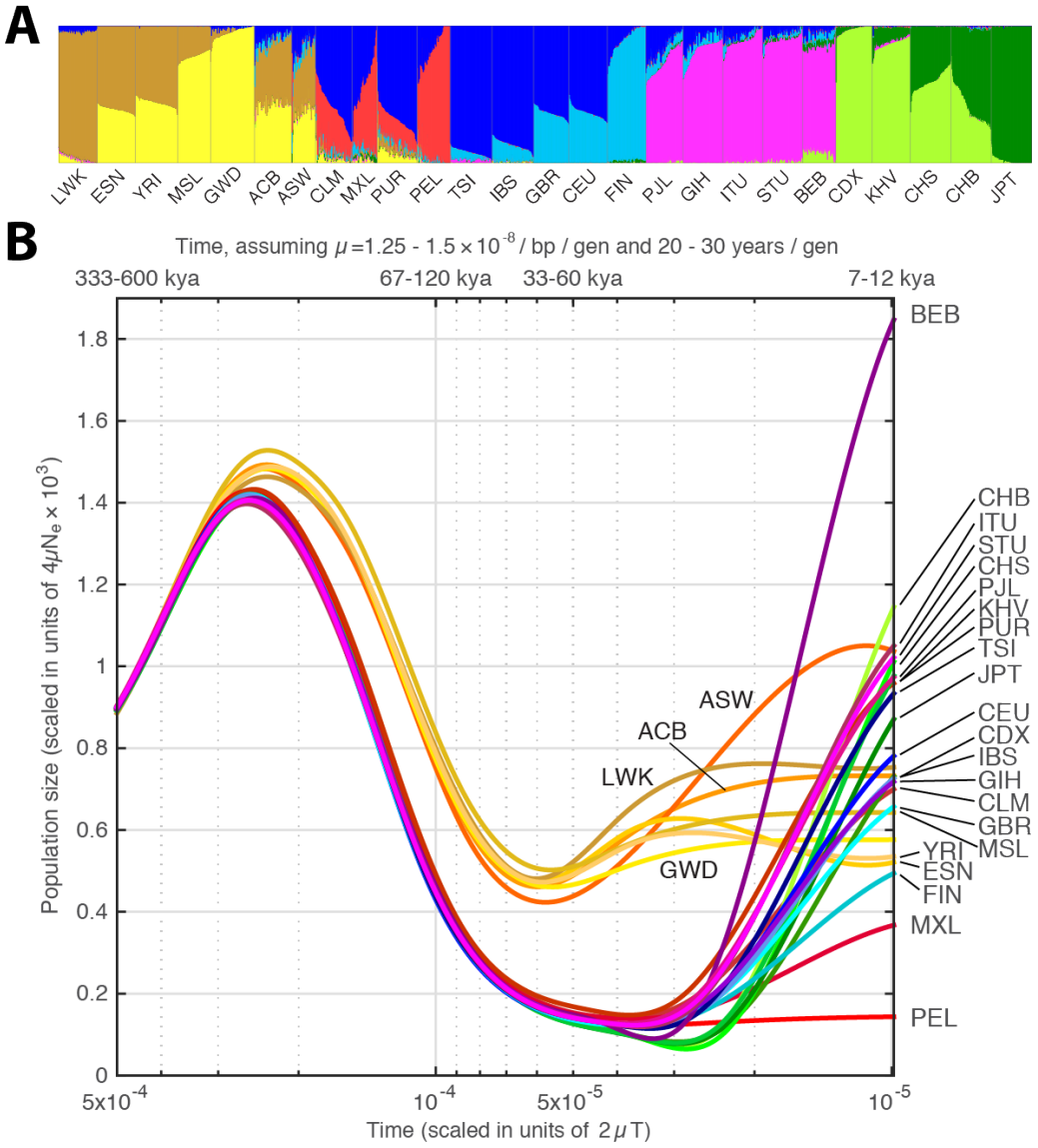


Figure 2: Population structure and demography. **A)** Population structure inferred using a maximum likelihood approach with 8 clusters. **B)** Changes to effective population sizes over time, inferred using PSMC. Lines represent the within-population median PSMC estimate, smoothed by fitting a cubic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

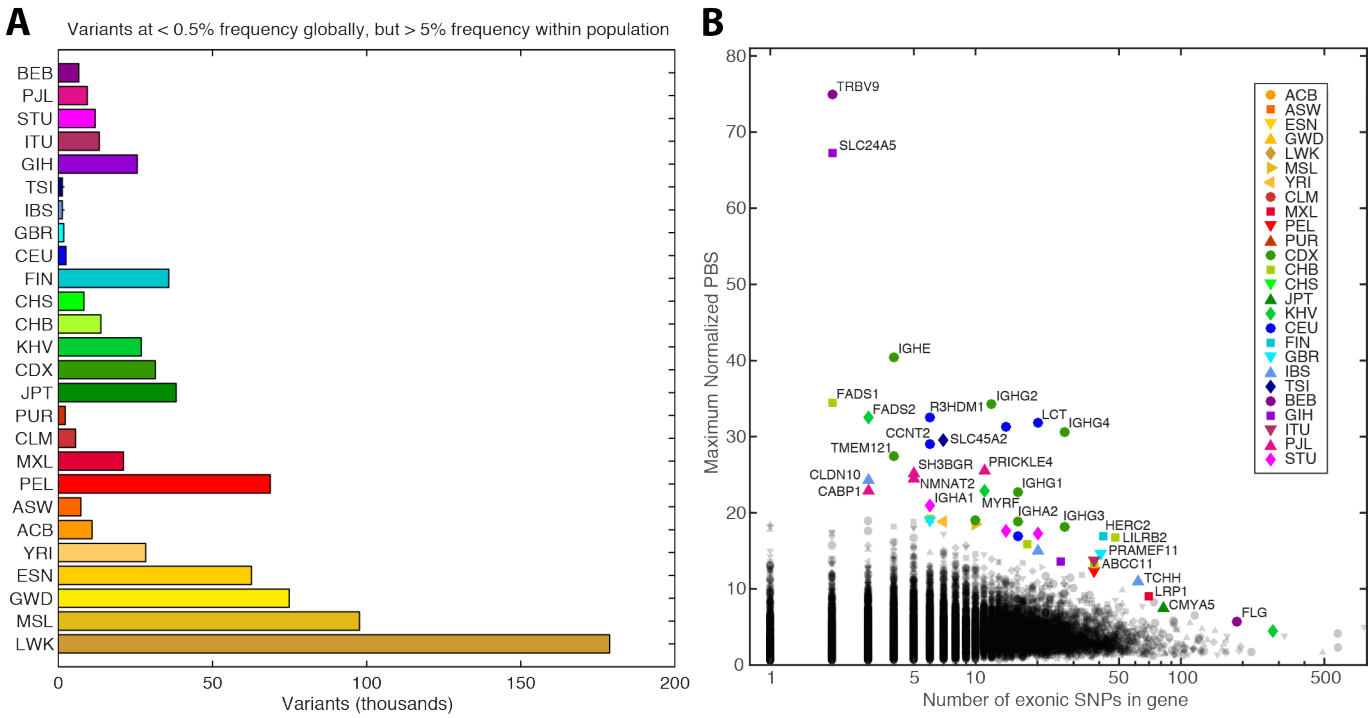


Figure 3: Population differentiation. **A)** Variants found to be rare (<0.5%) within the global sample, but common (>5%) within a population. **B)** Genes showing strong differentiation between pairs of closely related populations. The vertical axis gives the maximum obtained value of the F_{ST}-based population branch statistic (PBS), with selected genes colored to indicate the population in which the maximum value was achieved.

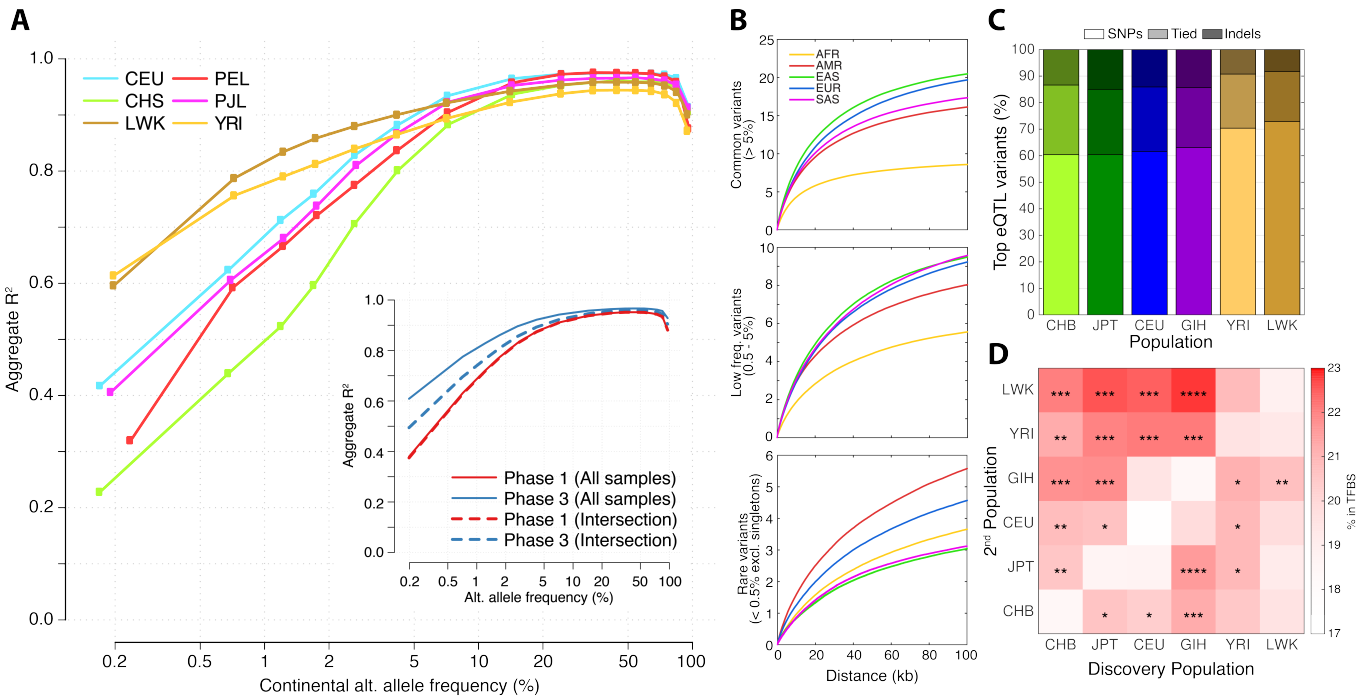
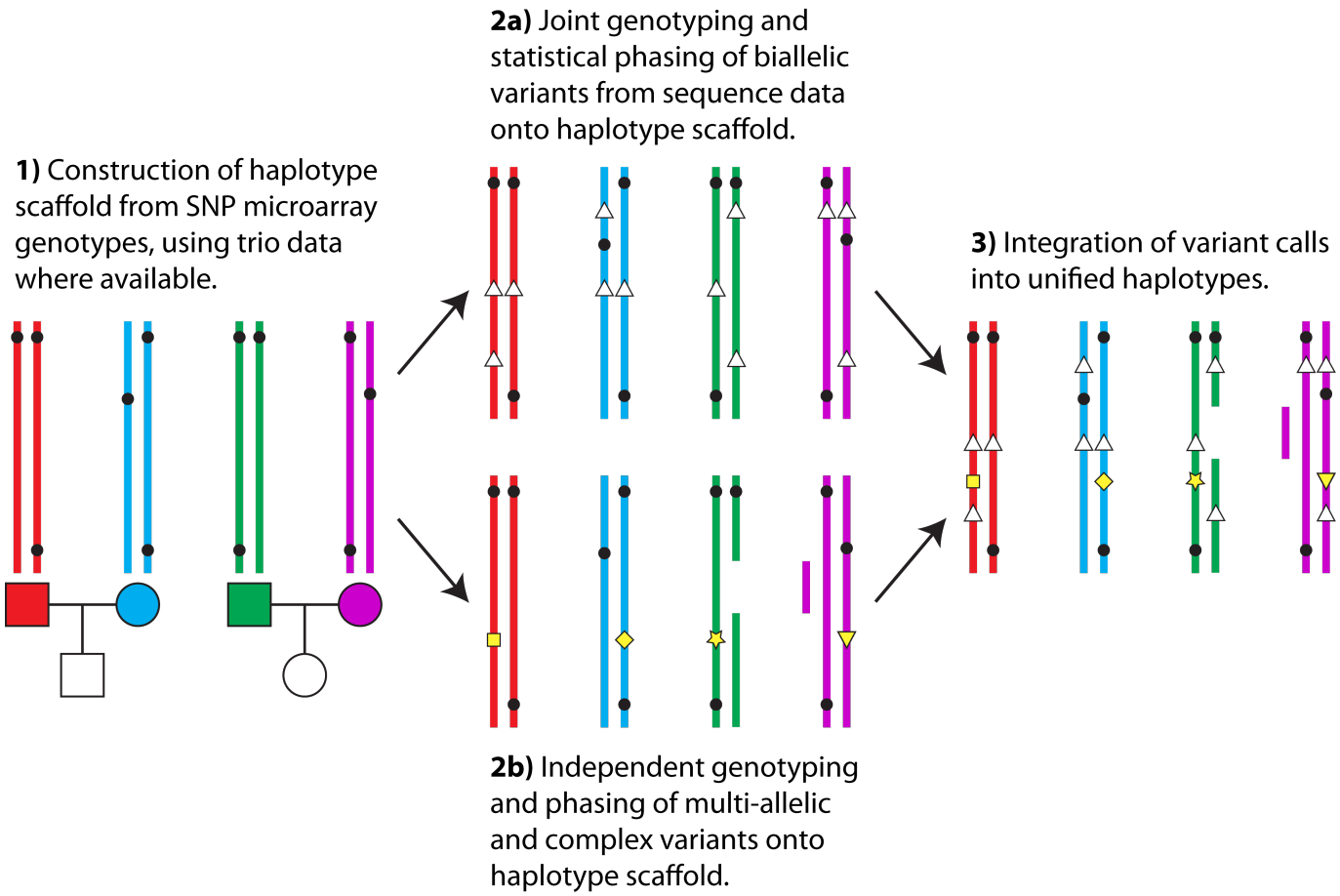


Figure 4: Imputation and eQTL discovery. **A)** Imputation accuracy as a function of allele frequency for six populations. The insert compares imputation accuracy between Phase 3 and Phase 1, using all samples (solid lines) and intersecting samples (dashed lines). **B)** The average number of tagging 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 (One star: p<0.01, two stars: p<0.001, three stars: p<0.0001, McNemar's test). The diagonal represents the percentage of eQTLs in TFBS using the original discovery sample.


2a) Joint genotyping and statistical phasing of biallelic variants from sequence data onto haplotype scaffold.
1) Construction of haplotype scaffold from SNP microarray genotypes, using trio data where available.
3) Integration of variant calls into unified haplotypes.
2b) Independent genotyping and phasing of multi-allelic and complex variants onto haplotype scaffold.

Box 1 figure.