ARTICLE

OPEN

doi:10.1038/nature15393

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 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) (Fig. 1a; see Supplementary Table 1 for population descriptions and abbreviations). All individuals were sequenced using both whole-genome sequencing (mean depth = 7.4×) and targeted exome sequencing (mean depth = 65.7×). 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].

Data set 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 (SVs). An overview of the sample collection, data generation, data processing, and analysis is given in Extended Data Fig. 1. 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 high confidence bi-allelic variants that were analysed 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 million of the 100 million variants in the public dbSNP catalogue (version 141 includes 40 million SNPs and indels newly contributed by this analysis). These novel variants especially enhance our catalogue 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 (>30×) 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 = 47×) 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% (Extended Data Fig. 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 threefold 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 4.1 million to 5.0 million sites (Fig. 1b and Table 1). Although >99.9% of variants consist of SNPs and short indels, structural variants affect more bases: the typical genome contains an estimated 2,100 to 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.

*Lists of participants and their affiliations appear in the online version of the paper.

©2015 Macmillan Publishers Limited. All rights reserved

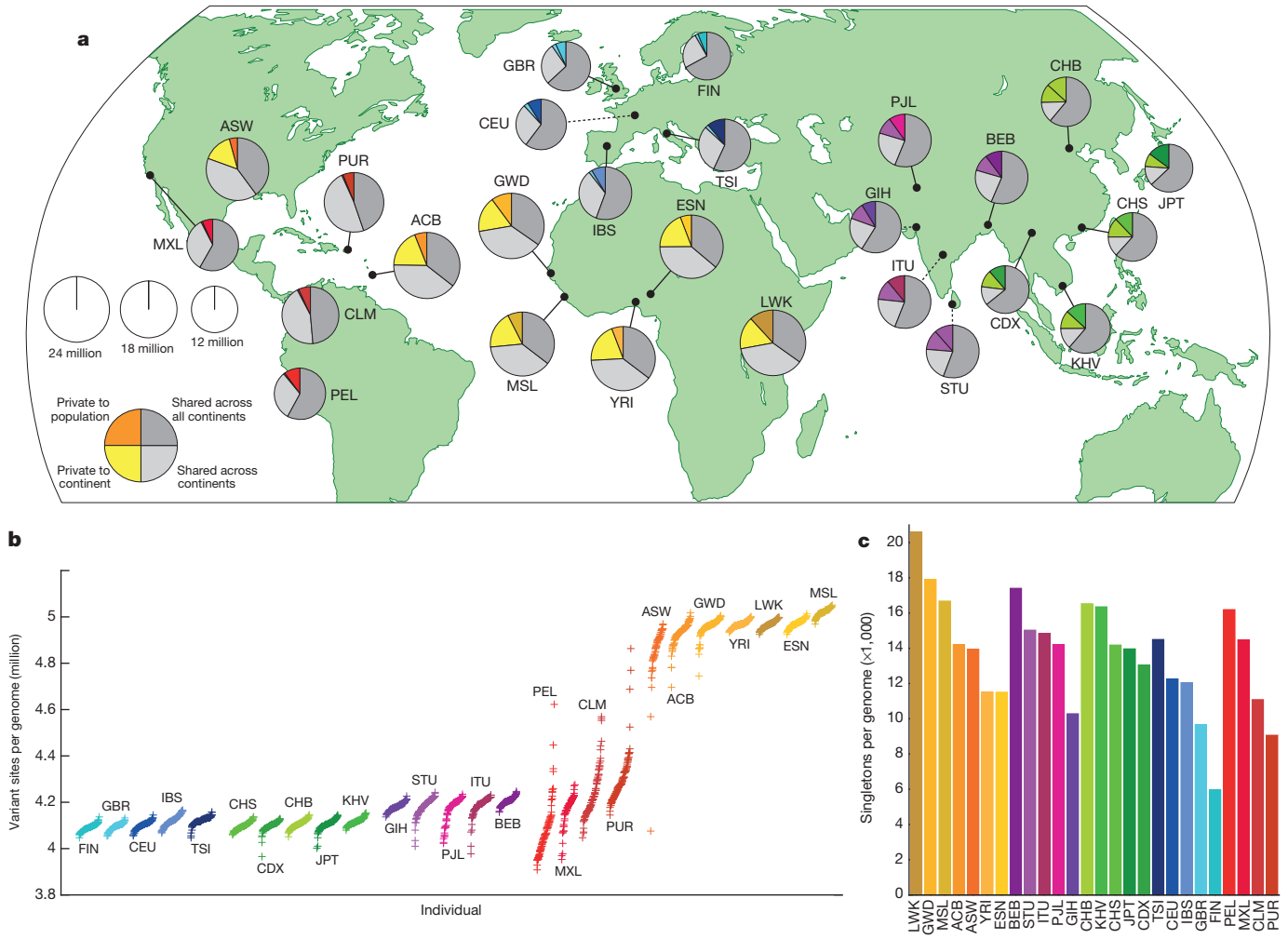


Figure 1 | Population sampling. **a**, Polymorphic variants within sampled populations. The area of each pie is proportional to the number of polymorphisms within a population. Pies are divided into four slices, representing variants private to a population (darker colour unique to population), private to a continental area (lighter colour 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.

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[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

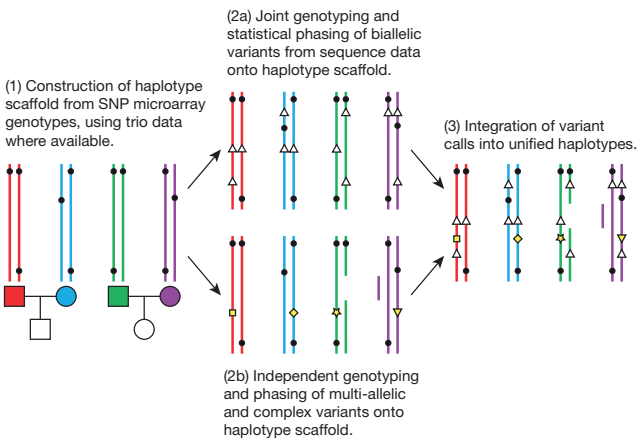
Systematic analysis of the patterns in which genetic variants are shared among individuals and populations provides detailed accounts of population history. Although most common variants are shared across the world, rarer variants are typically restricted to closely related populations (Fig. 1a); 86% of variants were restricted to a single continental group. Using a maximum likelihood approach[12], we estimated the proportion of each genome derived from several putative 'ancestral populations' (Fig. 2a and Extended Data Fig. 5).

©2015 Macmillan Publishers Limited. All rights reserved

BOX 1

Building a haplotype scaffold

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 (black circles) 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 ~1,062 kb, with typical phasing errors stretching ~37 kb (Supplementary Table 12).



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 (Extended Data Fig. 6a, b). Analysis of shared haplotype lengths around f_2 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 populations[14]. 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). 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 (Fig. 3b). This approach reveals a number of previously identified selection signals (such as *SLC24A5* associated with skin pigmentation[17], *HERC2* associated with eye colour[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; 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 evolution[20].

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

©2015 Macmillan Publishers Limited. All rights reserved

Table 1 | Median autosomal variant sites per genome

	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 (L1)	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
TFBSs	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

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.

most accurate in African ancestry populations, where greater genetic diversity results in a larger number of haplotypes and improves the chances that a rare variant is tagged by a characteristic haplotype.

Resolution of genetic association studies

To evaluate the impact of our new reference panel on GWAS, we re-analysed a previous study of age-related macular degeneration (AMD) totalling 2,157 cases and 1,150 controls[21]. We imputed 17.0 million genetic variants with estimated $R^2 > 0.3$, compared to 14.1 million variants using phase 1, and only 2.4 million SNPs using HapMap2. Compared to phase 1, the number of imputed common and intermediate frequency variants increased by 7%, whereas 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 false positives and false negatives[22–24]. We recommend that thresholds aiming for strict control of false positives should be determined using permutations. We expect thresholds to become more stringent when larger sample sizes are sequenced, when diverse samples are studied, or when genotyping and imputation is replaced with direct sequencing. 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 ~4× larger than in HapMap2-based analysis and 7% larger than in analyses based on phase 1 (Supplementary Table 9). In the *ARMS2/HTRA1* 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 (Extended Data Fig. 10). To evaluate how these differences influence the resolution of genetic association studies and,

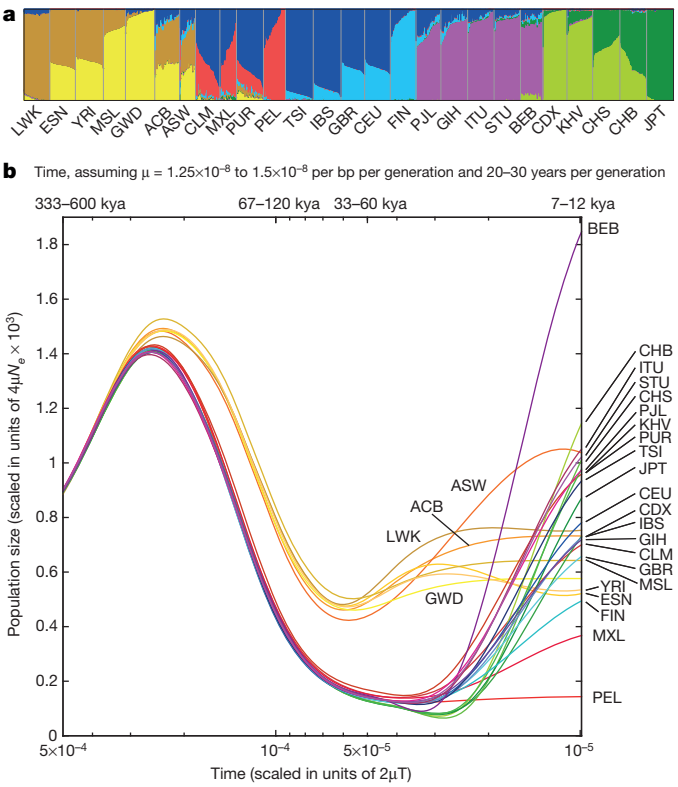


Figure 2 | Population structure and demography. **a**, Population structure inferred using a maximum likelihood approach with 8 clusters. **b**, Changes to effective population sizes over time, inferred using PSMC. Lines represent the within-population median PSMC estimate, smoothed by fitting a cubic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

©2015 Macmillan Publishers Limited. All rights reserved

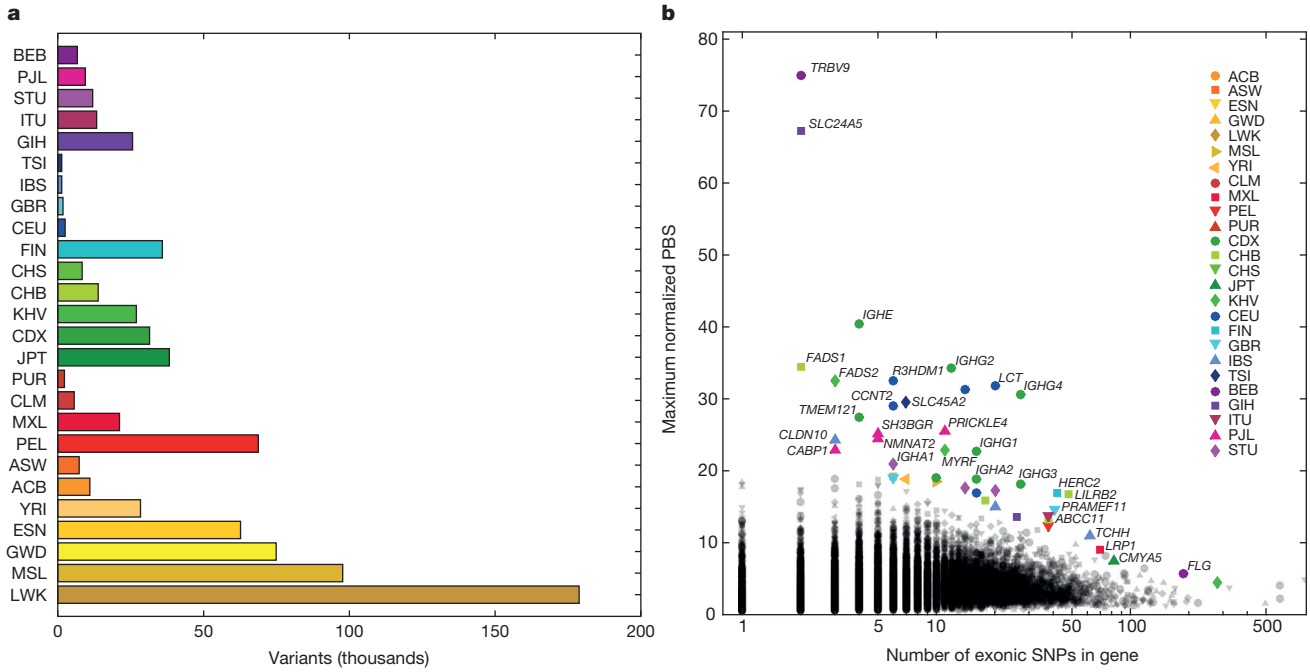


Figure 3 | Population differentiation. **a**, Variants found to be rare (<0.5%) within the global sample, but common (>5%) within a population. **b**, Genes showing strong differentiation between pairs of closely related populations. The vertical axis gives the maximum obtained value of the F_{ST}-based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

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

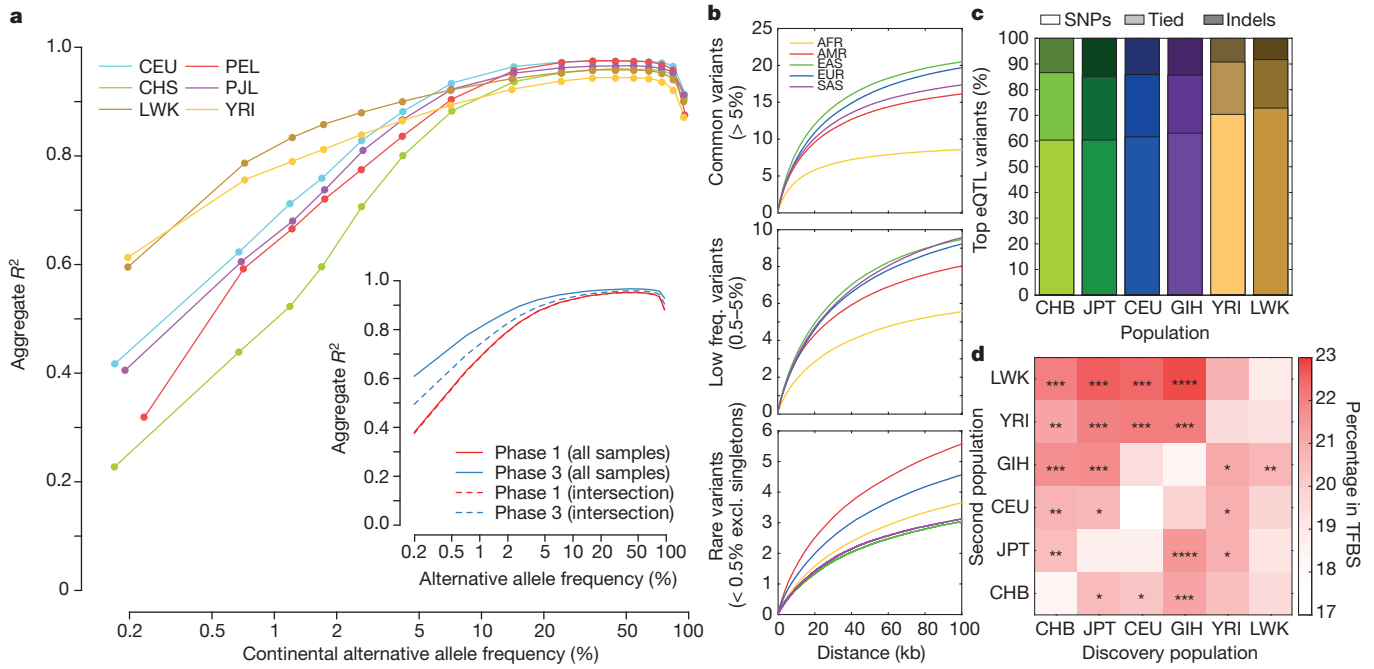


Figure 4 | Imputation and eQTL discovery. **a**, Imputation accuracy as a function of allele frequency for six populations. The insert compares imputation accuracy between phase 3 and phase 1, using all samples (solid lines) and intersecting samples (dashed lines). **b**, The average number of tagging 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.

©2015 Macmillan Publishers Limited. All rights reserved

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, 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 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[33]. 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[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. 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.

1. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
2. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
3. Voight, B. F. *et al.* The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet.* **8,** e1002793 (2012).
4. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nature Genet.* **43,** 1193–1201 (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 Genet.* **44,** 955–959 (2012).
6. Xue, Y. *et al.* Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *Am. J. Hum. Genet.* **91,** 1022–1032 (2012).
7. Jung, H., Bleazard, T., Lee, J. & Hong, D. Systematic investigation of cancer-associated somatic point mutations in SNP databases. *Nature Biotechnol.* **31,** 787–789 (2013).
8. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* http://dx.doi.org/10.1038/nature15394 (this issue).
9. The Haplotype Reference Consortium (http://www.haplotype-reference-consortium.org/).
10. Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nature Genet.* **46,** 220–224 (2014).
11. Do, R. *et al.* No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nature Genet.* **47,** 126–131 (2015).
12. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19,** 1655–1664 (2009).
13. Mathieson, I. & McVean, G. Demography and the age of rare variants. *PLoS Genet.* **10,** e1004528 (2014).
14. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475,** 493–496 (2011).
15. Moltke, I. *et al.* A common Greenlandic *TBC1D4* variant confers muscle insulin resistance and type 2 diabetes. *Nature* **512,** 190–193 (2014).
16. Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329,** 75–78 (2010).
17. Lamason, R. L. *et al.* SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* **310,** 1782–1786 (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. *Hum. Genet.* **123,** 177–187 (2008).
19. Mathias, R. A. *et al.* Adaptive evolution of the *FADS* gene cluster within Africa. *PLoS ONE* **7,** e44926 (2012).
20. Hernandez, R. D. *et al.* Classic selective sweeps were rare in recent human evolution. *Science* **331,** 920–924 (2011).
21. Chen, W. *et al.* Genetic variants near *TIMP3* and high-density lipoprotein-associated loci influence susceptibility to age-related macular degeneration. *Proc. Natl Acad. Sci. USA* **107,** 7401–7406 (2010).
22. Wakefield, J. Bayes factors for genome-wide association studies: comparison with *P*-values. *Genet. Epidemiol.* **33,** 79–86 (2009).
23. Wakefield, J. Commentary: genome-wide significance thresholds via Bayes factors. *Int. J. Epidemiol.* **41,** 286–291 (2012).
24. Sham, P. C. & Purcell, S. M. Statistical power and significance testing in large-scale genetic studies. *Nature Rev. Genet.* **15,** 335–346 (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 Genet.* **38,** 458–462 (2006).
26. Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308,** 385–389 (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 (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 (2007).
29. Maller, J. B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genet.* **44,** 1294–1301 (2012).
30. Fritsche, L. G. *et al.* Age-related macular degeneration is associated with an unstable *ARMS2* (*LOC387715*) mRNA. *Nature Genet.* **40,** 892–896 (2008).
31. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
32. Stranger, B. E. *et al.* Patterns of *cis* regulatory variation in diverse human populations. *PLoS Genet.* **8,** e1002639 (2012).
33. Chaisson, M. J. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517,** 608–611 (2015).

©2015 Macmillan Publishers Limited. All rights reserved

34. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nature Genet.* **47,** 435–444 (2015).
35. The UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* http://dx.doi.org/10.1038/nature14962 (2015).
36. Sidore, C. *et al.* Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nature Genet.* http://dx.doi.org/10.1038/ng3368 (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. *Nature Commun.* **5,** 3934 (2014).
38. O'Connell, J. *et al.* A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.* **10,** e1004234 (2014).
39. Menelaou, A. & Marchini, J. Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. *Bioinformatics* **29,** 84–91 (2013).

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; Esan 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 in Webuye, Kenya; Mende 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; Toscani in Italia; 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. Sadzewicz, J. Liu, and L. Tallon; 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. Drury and H. Arbery; A. Naranjo, M. Victoria Parra, and C. Duque; S. Dökel, B. Lenz, and S. Schrinner; S. Bumpstead; and C. Fletcher-Hoppe. Funding for this work was from the Wellcome Trust Core Award 090532/Z/09/Z and Senior Investigator Award 095552/Z/11/Z (P.D.), and grants WT098051 (R.D.), WT095908 and WT109497 (P.F.), WT086084/Z/08/Z and WT100956/Z/13/Z (G.M.), WT097307 (W.K.), WT0855322/Z/08/Z (R.L.), WT090770/Z/09/Z (D.K.), the Wellcome Trust Major Overseas program in Vietnam grant 089276/Z.09/Z (S.D.), the Medical Research Council UK grant G0801823 (J.L.M.), the UK Biotechnology and Biological Sciences Research Council grants BB/I02593X/1 (G.M.) and BB/I021213/1 (A.R.L.), the British Heart Foundation (C.A.A.), the Monument Trust (J.H.), the European Molecular Biology Laboratory (P.F.), the European Research Council grant 617306 (J.L.M.), the Chinese 863 Program 2012AA02A201, the National Basic Research program of China 973 program no. 2011CB809201, 2011CB809202 and 2011CB809203, Natural Science Foundation of China 31161130357, the Shenzhen Municipal Government of China grant ZYC201105170397A (J.W.), the Canadian Institutes of Health Research Operating grant 136855 and Canada Research Chair (S.G.), Banting Postdoctoral Fellowship from the Canadian Institutes of Health Research (M.K.D.), a Le Fonds de Recherche du Québec-Santé (FRQS) research fellowship (A.H.), Genome Quebec (P.A.), the Ontario Ministry of Research and Innovation – Ontario Institute for Cancer Research Investigator Award (P.A., J.S.), the Quebec Ministry of Economic Development, Innovation, and Exports grant PSR-SIIRI-195 (P.A.), the German Federal Ministry of Education and Research (BMBF) grants 0315428A and 01GS08201 (R.H.), the Max Planck Society (H.L., G.M., R.S.), BMBF-EPITREAT grant 0316190A (R.H., M.L.), the German Research Foundation (Deutsche Forschungsgemeinschaft) Emmy Noether Grant KO4037/1-1 (J.O.K.), the Beatriu de Pinos Program grants 2006 BP-A 10144 and 2009 BP-B 00274 (M.V.), the Spanish National Institute for Health Research grant PRB2 IPT13/0001-ISCIII-SGEFI/FEDER (A.O.), Ewha Womans University (C.L.), the Japan Society for the Promotion of Science Fellowship number PE13075 (N.P.), the Louis Jeantet Foundation (E.T.D.), the Marie Curie Actions Career Integration grant 303772 (C.A.), the Swiss National Science Foundation 31003A_130342 and NCCR "Frontiers in Genetics" (E.T.D.), the University of Geneva (E.T.D., T.L., G.M.), the US National Institutes of Health National Center for Biotechnology Information (S.S.) and grants U54HG3067 (E.S.L.), U54HG3273 and U01HG5211 (R.A.G.), U54HG3079 (R.K.W., E.R.M.), R01HG2898 (S.E.D.), R01HG2385 (E.E.E.), RC2HG5552 and U01HG6513 (G.T.M., G.R.A.), U01HG5214 (A.C.), U01HG5715 (C.D.B.), U01HG5718 (M.G.), U01HG5728 (Y.X.F.), U41HG7635 (R.K.W., E.E.E., P.H.S.), U41HG7497 (C.L., M.A.B., K.C., L.D., E.E.E., M.G., J.O.K., G.T.M., S.A.M., R.E.M., J.L.S., K.Y.), R01HG4960 and R01HG5701 (B.L.B.), R01HG5214 (G.A.), R01HG6855 (S.M.), R01HG7068 (R.E.M.), R01HG7644 (R.D.H.), DP2OD6514 (P.S.), DP5OD9154 (J.K.), R01CA166661 (S.E.D.), R01CA172652 (K.C.), P01GM99568 (S.R.B.), R01GM59290 (L.B.J., M.A.B.), R01GM104390 (L.B.J., M.Y.Y.), T32GM7790 (C.D.B., A.R.M.), P01GM99568 (S.R.B.), R01HL87699 and R01HL104608 (K.C.B.), T32HL94284 (J.L.R.F.), and contracts HHSN268201100040C (A.M.R.) and HHSN272201000025C (P.S.), Harvard Medical School Eleanor and Miles Shore Fellowship (K.L.), Lundbeck Foundation Grant R170-2014-1039 (K.L.), NIJ Grant 2014-DN-BX-K089 (Y.E.), the Mary Beryl Patch Turnbull Scholar Program (K.C.B.), NSF Graduate Research Fellowship DGE-1147470 (G.D.P.), the Simons Foundation SFARI award SF51 (M.W.), and a Sloan Foundation Fellowship (R.D.H.). E.E.E. is an investigator of the Howard Hughes Medical Institute.

Author Contributions 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

©2015 Macmillan Publishers Limited. All rights reserved

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

Corresponding authors Adam Auton[1], Gonçalo R. Abecasis[2]

Steering committee: David M. Altshuler[3] (Co-Chair), Richard M. Durbin[4] (Co-Chair), Gonçalo R. Abecasis[2], David R. Bentley[5], Aravinda Chakravarti[6], Andrew G. Clark[7], Peter Donnelly[8,9], Evan E. Eichler[10,11], Paul Flicek[12], Stacey B. Gabriel[13], Richard A. Gibbs[14], Eric D. Green[15], Matthew E. Hurles[4], Bartha M. Knoppers[16], Jan O. Korbel[12,17], Eric S. Lander[13], Charles Lee[18,19], Hans Lehrach[20,21], Elaine R. Mardis[22], Gabor T. Marth[23], Gil A. McVean[8,9], Deborah A. Nickerson[10], Jeanette P. Schmidt[24], Stephen T. Sherry[25], Jun Wang[26,27,28,29,30], Richard K. Wilson[22]

Production group: Baylor College of Medicine Richard A. Gibbs[14] (Principal Investigator), Eric Boerwinkle[14], Harsha Doddapaneni[14], Yi Han[14], Viktoriya Korchina[14], Christie Kovar[14], Sandra Lee[14], Donna Muzny[14], Jeffrey G. Reid[14], Yiming Zhu[14]; **BGI-Shenzhen** Jun Wang[26,27,28,29,30] (Principal Investigator), Yuqi Chang[26], Qiang Feng[26,27], Xiaodong Fang[26,27], Xiaosen Guo[26,27], Min Jian[26,27], Hui Jiang[26,27], Xin Jin[26], Tianming Lan[26], Guoqing Li[26], Jingxiang Li[26], Yingrui Li[26], Shengmao Liu[26], Xiao Liu[26,27], Yao Lu[26], Xuedi Ma[26], Meifang Tang[26], Bo Wang[26], Guangbiao Wang[26], Honglong Wu[26], Renhua Wu[26], Xun Xu[26], Ye Yin[26], Dandan Zhang[26], Wenwei Zhang[26], Jiao Zhao[26], Meiru Zhao[26], Xiaole Zheng[26]; **Broad Institute of MIT and Harvard** Eric S. Lander[13] (Principal Investigator), David M. Altshuler[3], Stacey B. Gabriel[13] (Co-Chair), Namrata Gupta[13]; **Coriell Institute for Medical Research** Neda Gharani[31], Lorraine H. Toji[31], Norman P. Gerry[31], Alissa M. Resch[31]; **European Molecular Biology Laboratory, European Bioinformatics Institute** Paul Flicek[12] (Principal Investigator), Jonathan Barker[12], Laura Clarke[12], Laurent Gil[12], Sarah E. Hunt[12], Gavin Kelman[12], Eugene Kulesha[12], Rasko Leinonen[12], William M. McLaren[12], Rajesh Radhakrishnan[12], Asier Roa[12], Dmitriy Smirnov[12], Richard E. Smith[12], Ian Streeter[12], Anja Thormann[12], Iliana Toneva[12], Brendan Vaughan[12], Xiangqun Zheng-Bradley[12]; **Illumina** David R. Bentley[5] (Principal Investigator), Russell Grocock[5], Sean Humphray[5], Terena James[5], Zoya Kingsbury[5]; **Max Planck Institute for Molecular Genetics** Hans Lehrach[20,21] (Principal Investigator), Ralf Sudbrak[32](Project Leader), Marcus W. Albrecht[33], Vyacheslav S. Amstislavskiy[20], Tatiana A. Borodina[33], Matthias Lienhard[20], Florian Mertes[20], Marc Sultan[20], Bernd Timmermann[20], Marie-Laure Yaspo[20]; **McDonnell Genome Institute at Washington University** Elaine R. Mardis[22] (Co-Principal Investigator) (Co-Chair), Richard K. Wilson[22] (Co-Principal Investigator), Lucinda Fulton[22], Robert Fulton[22]; **US National Institutes of Health** Stephen T. Sherry[25] (Principal Investigator), Victor Ananiev[25], Zinaida Belaia[25], Dimitriy Beloslyudtsev[25], Nathan Bouk[25], Chao Chen[25], Deanna Church[34], Robert Cohen[25], Charles Cook[25], John Garner[25], Timothy Hefferon[25], Mikhail Kimelman[25], Chunlei Liu[25], John Lopez[25], Peter Meric[25], Chris O'Sullivan[35], Yuri Ostapchuk[25], Lon Phan[25], Sergiy Ponomarov[25], Valerie Schneider[25], Eugene Shekhtman[25], Karl Sirotkin[25], Douglas Slotta[25], Hua Zhang[25]; **University of Oxford** Gil A. McVean[8,9] (Principal Investigator); **Wellcome Trust Sanger Institute** Richard M. Durbin[4] (Principal Investigator), Senduran Balasubramaniam[4], John Burton[4], Petr Danecek[4], Thomas M. Keane[4], Anja Kolb-Kokocinski[4], Shane McCarthy[4], James Stalker[4], Michael Quail[4]

Analysis group: Affymetrix Jeanette P. Schmidt[24] (Principal Investigator), Christopher J. Davies[24], Jeremy Gollub[24], Teresa Webster[24], Brant Wong[24], Yiping Zhan[24]; **Albert Einstein College of Medicine** Adam Auton[1] (Principal Investigator), Christopher L. Campbell[1], Yu Kong[1], Anthony Marcketta[1]; **Baylor College of Medicine** Richard A. Gibbs[14] (Principal Investigator), Fuli Yu[14] (Project Leader), Lilian Antunes[14], Matthew Bainbridge[14], Donna Muzny[14], Aniko Sabo[14], Zhuoyi Huang[14]; **BGI-Shenzhen** Jun Wang[26,27,28,29,30] (Principal Investigator), Lachlan J. M. Coin[26], Lin Fang[26,27], Xiaosen Guo[26], Xin Jin[26], Guoqing Li[26], Qibin Li[26], Yingrui Li[26], Zhenyu Li[26], Haoxiang Lin[26], Binghang Liu[26], Ruibang Luo[26], Haojing Shao[26], Yinlong Xie[26], Chen Ye[26], Chang Yu[26], Fan Zhang[26], Hancheng Zheng[26], Hongmei Zhu[26]; **Bilkent University** Can Alkan[36], Elif Dal[36], Fatma Kahveci[36]; **Boston College** Gabor T. Marth[23] (Principal Investigator), Erik P. Garrison[4] (Project Lead), Deniz Kural[37], Wan-Ping Lee[37], Wen Fung Leong[38], Michael Stromberg[39], Alistair N. Ward[23], Jiantao Wu[39], Mengyao Zhang[40]; **Broad Institute of MIT and Harvard** Mark J. Daly[13] (Principal Investigator), Mark A. DePristo[41] (Project Leader), Robert E. Handsaker[13,40] (Project Leader), David M. Altshuler[3], Eric Banks[13], Gaurav Bhatia[13], Guillermo del Angel[13], Stacey B. Gabriel[13], Giulio Genovese[13], Namrata Gupta[13], Heng Li[13], Seva Kashin[13,40], Eric S. Lander[13], Steven A. McCarroll[13,40], James C. Nemesh[13], Ryan E. Poplin[13]; **Cold Spring Harbor Laboratory** Seungtai C. Yoon[42] (Principal Investigator), Jayon Lihm[42], Vladimir Makarov[43]; **Cornell University** Andrew G. Clark[7] (Principal Investigator), Srikanth Gottipati[44], Alon Keinan[7], Juan L. Rodriguez-Flores[45]; **European Molecular Biology Laboratory** Jan O. Korbel[12,17] (Principal Investigator), Tobias Rausch[17,46] (Project Leader), Markus H. Fritz[46], Adrian M. Stütz[17]; **European Molecular Biology Laboratory, European Bioinformatics Institute** Paul Flicek[12] (Principal Investigator), Kathryn Beal[12], Laura Clarke[12], Avik Datta[12], Javier Herrero[47], William M. McLaren[12], Graham R. S. Ritchie[12], Richard E. Smith[12], Daniel Zerbino[12], Xiangqun Zheng-Bradley[12]; **Harvard University** Pardis C. Sabeti[13,48] (Principal Investigator), Ilya Shlyakhter[13,48], Stephen F. Schaffner[13,48], Joseph Vitti[13,49]; **Human Gene Mutation Database** David N. Cooper[50] (Principal Investigator), Edward V. Ball[50], Peter D. Stenson[50]; **Illumina** David R. Bentley[5] (Principal Investigator), Bret Barnes[39], Markus Bauer[5], R. Keira Cheetham[5], Anthony Cox[5], Michael Eberle[5], Sean Humphray[5], Scott Kahn[39], Lisa Murray[5], John Peden[5], Richard Shaw[5]; **Icahn School of Medicine at Mount Sinai** Eimear E. Kenny[51] (Principal Investigator); **Louisiana State University** Mark A. Batzer[52] (Principal Investigator), Miriam K. Konkel[52], Jerilyn A. Walker[52]; **Massachusetts General Hospital** Daniel G. MacArthur[53] (Principal Investigator), Monkol Lek[53]; **Max Planck Institute for Molecular Genetics** Ralf Sudbrak[32] (Project Leader), Vyacheslav S. Amstislavskiy[20], Ralf Herwig[20]; **McDonnell Genome Institute at Washington University** Elaine R. Mardis[22] (Co-Principal Investigator), Li Ding[22], Daniel C. Koboldt[22], David Larson[22], Kai Ye[22]; **McGill University** Simon Gravel[54]; **National Eye Institute, NIH** Anand Swaroop[55], Emily Chew[55]; **New York Genome Center** Tuuli Lappalainen[56,57] (Principal Investigator), Yaniv Erlich[56,58] (Principal Investigator), Melissa Gymrek[13,56,59,60], Thomas Frederick Willems[61]; **Ontario Institute for Cancer Research** Jared T. Simpson[62]; **Pennsylvania State University** Mark D. Shriver[63] (Principal Investigator); **Rutgers Cancer Institute of New Jersey** Jeffrey A. Rosenfeld[64] (Principal Investigator); **Stanford University** Carlos D. Bustamante[65] (Principal Investigator), Stephen B. Montgomery[66] (Principal Investigator), Francisco M. De La Vega[65] (Principal Investigator), Jake K. Byrnes[67], Andrew W. Carroll[68], Marianne K. DeGorter[66], Phil Lacroute[65], Brian K. Maples[65], Alicia R. Martin[65], Andres Moreno-Estrada[65,69], Suyash S. Shringarpure[65], Fouad Zakharia[65]; **Tel-Aviv University** Eran Halperin[70,71,72] (Principal Investigator), Yael Baran[70]; **The Jackson Laboratory for Genomic Medicine** Charles Lee[18,19] (Principal Investigator), Eliza Cerveira[18], Jaeho Hwang[18], Ankit Malhotra[18] (Co-Project Lead), Dariusz Plewczynski[18], Kamen Radew[18], Mallory Romanovitch[18], Chengsheng Zhang[18] (Co-Project Lead); **Thermo Fisher Scientific** Fiona C. L. Hyland[73]; **Translational Genomics Research Institute** David W. Craig[74] (Principal Investigator), Alexis Christoforides[74], Nils Homer[75], Tyler Izatt[74], Ahmet A. Kurdoglu[74], Shripad A. Sinari[74], Kevin Squire[76]; **US National Institutes of Health** Stephen T. Sherry[25] (Principal Investigator), Chunlin Xiao[25]; **University of California, San Diego** Jonathan Sebat[77,78] (Principal Investigator), Danny Antaki[77], Madhusudan Gujral[77], Amina Noor[77], Kenny Ye[79]; **University of California, San Francisco** Esteban G. Burchard[80] (Principal Investigator), Ryan D. Hernandez[80,81,82] (Principal Investigator), Christopher R. Gignoux[80]; **University of California, Santa Cruz** David Haussler[83,84] (Principal Investigator), Sol J. Katzman[83], W. James Kent[83]; **University of Chicago** Bryan Howie[85]; **University College London** Andres Ruiz-Linares[86] (Principal Investigator); **University of Geneva** Emmanouil T. Dermitzakis[87,88,89] (Principal Investigator); **University of Maryland School of Medicine** Scott E. Devine[90] (Principal Investigator); **University of Michigan** Gonçalo R. Abecasis[2] (Principal Investigator) (Co-Chair), Hyun Min Kang[2] (Project Leader), Jeffrey M. Kidd[91,92] (Principal Investigator), Tom Blackwell[2], Sean Caron[2], Wei Chen[93], Sarah Emery[92], Lars Fritsche[2], Christian Fuchsberger[2], Goo Jun[2,94], Bingshan Li[95], Robert Lyons[96], Chris Scheller[2], Carlo Sidore[2,97,98], Shiya Song[91], Elzbieta Sliwerska[92], Daniel Taliun[2], Adrian Tan[2], Ryan Welch[2], Mary Kate Wing[2], Xiaowei Zhan[99]; **University of Montréal** Philip Awadalla[62,100] (Principal Investigator), Alan Hodgkinson[100]; **University of North Carolina at Chapel Hill** Yun Li[101]; **University of North Carolina at Charlotte** Xinghua Shi[102] (Principal Investigator), Andrew Quitadamo[102]; **University of Oxford** Gerton Lunter[8] (Principal Investigator), Gil A. McVean[8,9] (Principal Investigator) (Co-Chair), Jonathan L. Marchini[8,9] (Principal Investigator), Simon Myers[8,9] (Principal Investigator), Claire Churchhouse[9], Olivier Delaneau[9,87], Anjali Gupta-Hinch[8], Warren Kretzschmar[8], Zamin Iqbal[8], Iain Mathieson[8], Androniki Menelaou[9,103], Andy Rimmer[87], Dionysia K. Xifara[8,9]; **University of Puerto Rico** Taras K. Oleksyk[104] (Principal Investigator); **University of Texas Health Sciences Center at Houston** Yunxin Fu[94] (Principal Investigator), Xiaoming Liu[94], Momiao Xiong[94]; **University of Utah** Lynn Jorde[105] (Principal Investigator), David Witherspoon[105], Jinchuan Xing[106]; **University of Washington** Evan E. Eichler[10,11] (Principal Investigator), Brian L. Browning[107] (Principal Investigator), Sharon R. Browning[108] (Principal Investigator), Fereydoun Hormozdiari[10], Peter H. Sudmant[10]; **Weill Cornell Medical College**, Ekta Khurana[109] (Principal Investigator); **Wellcome Trust Sanger Institute** Richard M. Durbin[4] (Principal Investigator), Matthew E. Hurles[4] (Principal Investigator), Chris Tyler-Smith[4] (Principal Investigator), Cornelis A. Albers[110,111], Qasim Ayub[4], Senduran Balasubramaniam[4], Yuan Chen[4], Vincenza Colonna[4,112], Petr Danecek[4], Luke Jostins[8], Thomas M. Keane[4], Shane McCarthy[4], Klaudia Walter[4], Yali Xue[4]; **Yale University** Mark B. Gerstein[113,114,115] (Principal Investigator), Alexej Abyzov[116], Suganthi Balasubramanian[115], Jieming Chen[113], Declan Clarke[117], Yao Fu[113], Arif O. Harmanci[113], Mike Jin[115], Donghoon Lee[113], Jeremy Liu[115], Xinmeng Jasmine Mu[13,113], Jing Zhang[113,115], Yan Zhang[113,115]

Structural variation group: BGI-Shenzhen Yingrui Li[26], Ruibang Luo[26], Hongmei Zhu[26]; **Bilkent University** Can Alkan[36], Elif Dal[36], Fatma Kahveci[36]; **Boston College** Gabor T. Marth[23] (Principal Investigator), Erik P. Garrison[4], Deniz Kural[37], Wan-Ping Lee[37], Alistair N. Ward[23], Jiantao Wu[23], Mengyao Zhang[23]; **Broad Institute of MIT and Harvard** Steven A. McCarroll[13,40] (Principal Investigator), Robert E. Handsaker[13,40] (Project Leader), David M. Altshuler[3], Eric Banks[13], Guillermo del Angel[13], Giulio Genovese[13], Chris Hartl[13], Heng Li[13], Seva Kashin[13,40], James C. Nemesh[13], Khalid Shakir[13]; **Cold Spring Harbor Laboratory** Seungtai C. Yoon[42] (Principal Investigator), Jayon Lihm[42], Vladimir Makarov[43]; **Cornell University** Jeremiah Degenhardt[7]; **European Molecular Biology Laboratory** Jan O. Korbel[12,17] (Principal Investigator) (Co-Chair), Markus H. Fritz[46], Sascha Meiers[17], Benjamin Raeder[17], Tobias Rausch[17,46], Adrian M. Stütz[17]; **European Molecular Biology Laboratory, European Bioinformatics Institute** Paul Flicek[12] (Principal Investigator), Francesco Paolo Casale[12], Laura Clarke[12], Richard E. Smith[12], Oliver Stegle[12], Xiangqun Zheng-Bradley[12]; **Illumina** David R. Bentley[5] (Principal Investigator), Bret Barnes[39], R. Keira Cheetham[5], Michael Eberle[5], Sean Humphray[5], Scott Kahn[39], Lisa Murray[5], Richard Shaw[5]; **Leiden University Medical Center** Eric-Wubbo Lameijer[118]; **Louisiana State University** Mark A. Batzer[52] (Principal Investigator), Miriam K. Konkel[52], Jerilyn A. Walker[52]; **McDonnell Genome Institute at Washington University** Li Ding[22] (Principal Investigator), Ira Hall[22], Kai Ye[22]; **Stanford University** Phil Lacroute[65]; **The Jackson Laboratory for Genomic Medicine** Charles Lee[18,19] (Principal Investigator) (Co-Chair), Eliza Cerveira[18], Ankit Malhotra[18], Jaeho Hwang[18], Dariusz Plewczynski[18], Kamen Radew[18], Mallory Romanovitch[18], Chengsheng Zhang[18]; **Translational Genomics Research Institute** David W. Craig[74] (Principal Investigator), Nils Homer[75]; **US National Institutes of Health** Deanna Church[34], Chunlin Xiao[25]; **University of California, San Diego** Jonathan Sebat[77] (Principal Investigator), Danny Antaki[77], Vineet Bafna[119], Jacob Michaelson[120], Kenny Ye[79]; **University of Maryland School of Medicine** Scott E. Devine[90] (Principal Investigator), Eugene J. Gardner[90] (Project Leader); **University of Michigan** Gonçalo R. Abecasis[2] (Principal Investigator), Jeffrey M. Kidd[91,92] (Principal Investigator), Ryan E. Mills[91,92] (Principal Investigator), Gargi

©2015 Macmillan Publishers Limited. All rights reserved

Dayama[91,92], Sarah Emery[92], Goo Jun[2,94]; **University of North Carolina at Charlotte** Xinghua Shi[102] (Principal Investigator), Andrew Quitadamo[102]; **University of Oxford** Gerton Lunter[8] (Principal Investigator), Gil A. McVean[8,9] (Principal Investigator); **University of Texas MD Anderson Cancer Center** Ken Chen[121] (Principle Investigator), Xian Fan[121], Zechen Chong[121], Tenghui Chen[121]; **University of Utah** David Witherspoon[105]; Jinchuan Xing[106]; **University of Washington** Evan E. Eichler[10,11] (Principal Investigator) (Co-Chair), Mark J. Chaisson[10], Fereydoun Hormozdiari[10], John Huddleston[10,11], Maika Malig[10], Bradley J. Nelson[10], Peter H. Sudmant[10]; **Vanderbilt University School of Medicine** Nicholas F. Parrish[95]; **Weill Cornell Medical College** Ekta Khurana[109] (Principal Investigator); **Wellcome Trust Sanger Institute** Matthew E. Hurles[4] (Principal Investigator), Ben Blackburne[4], Sarah J. Lindsay[4], Zemin Ning[4], Klaudia Walter[4], Yujun Zhang[4]; **Yale University** Mark B. Gerstein[113,114,115] (Principal Investigator), Alexej Abyzov[116], Jieming Chen[113], Declan Clarke[117], Hugo Lam[122], Xinmeng Jasmine Mu[13,113], Cristina Sisu[113], Jing Zhang[113,115], Yan Zhang[113,115]

Exome group: Baylor College of Medicine Richard A. Gibbs[14] (Principal Investigator) (Co-Chair), Fuli Yu[14] (Project Leader), Matthew Bainbridge[14], Danny Challis[14], Uday S. Evani[14], Christie Kovar[14], James Lu[14], Donna Muzny[14], Uma Nagaswamy[14], Jeffrey G. Reid[14], Aniko Sabo[14], Jin Yu[14]; **BGI-Shenzhen** Xiaosen Guo[26,27], Wangshen Li[26], Yingrui Li[26], Renhua Wu[26]; **Boston College** Gabor T. Marth[23] (Principal Investigator) (Co-Chair), Erik P. Garrison[4], Wen Fung Leong[23], Alistair N. Ward[23]; **Broad Institute of MIT and Harvard** Guillermo del Angel[13], Mark A. DePristo[41], Stacey B. Gabriel[13], Namrata Gupta[13], Chris Hartl[13], Ryan E. Poplin[13]; **Cornell University** Andrew G. Clark[7] (Principal Investigator), Juan L. Rodriguez-Flores[45]; **European Molecular Biology Laboratory, European Bioinformatics Institute** Paul Flicek[12] (Principal Investigator), Laura Clarke[12], Richard E. Smith[12], Xiangqun Zheng-Bradley[12]; **Massachusetts General Hospital** Daniel G. MacArthur[53] (Principal Investigator); **McDonnell Genome Institute at Washington University** Elaine R. Mardis[22] (Principal Investigator); Robert Fulton[22], Daniel C. Koboldt[22]; **McGill University** Simon Gravel[54]; **Stanford University** Carlos D. Bustamante[65] (Principal Investigator); **Translational Genomics Research Institute** David W. Craig[74] (Principal Investigator), Alexis Christoforides[74], Nils Homer[75], Tyler Izatt[74]; **US National Institutes of Health** Stephen T. Sherry[25] (Principal Investigator), Chunlin Xiao[25]; **University of Geneva** Emmanouil T. Dermitzakis[87,88,89] (Principal Investigator); **University of Michigan** Gonçalo R. Abecasis[2] (Principal Investigator), Hyun Min Kang[2]; **University of Oxford** Gil A. McVean[8,9] (Principal Investigator); **Yale University** Mark B. Gerstein[113,114,115] (Principal Investigator), Suganthi Balasubramanian[115], Lukas Habegger[113]

Functional interpretation group: Cornell University Haiyuan Yu[44] (Principal Investigator); **European Molecular Biology Laboratory, European Bioinformatics Institute** Paul Flicek[12] (Principal Investigator), Laura Clarke[12], Fiona Cunningham[12], Ian Dunham[12], Daniel Zerbino[12], Xiangqun Zheng-Bradley[12]; **Harvard University** Kasper Lage[13,123] (Principal Investigator), Jakob Berg Jespersen[13,123,124], Heiko Horn[13,123]; **Stanford University** Stephen B. Montgomery[66] (Principal Investigator), Marianne K. DeGorter[66]; **Weill Cornell Medical College**, Ekta Khurana[109] (Principal Investigator); **Wellcome Trust Sanger Institute** Chris Tyler-Smith[4] (Principal Investigator) (Co-Chair), Yuan Chen[4], Vincenza Colonna[4,112], Yali Xue[4]; **Yale University** Mark B. Gerstein[113,1114,115] (Principal Investigator) (Co-Chair), Suganthi Balasubramanian[115], Yao Fu[113], Donghoon Kim[115]

Chromosome Y group: Albert Einstein College of Medicine Adam Auton[1] (Principal Investigator), Anthony Marcketta[1]; **American Museum of Natural History** Rob Desalle[125], Apurva Narechania[126]; **Arizona State University** Melissa A. Wilson Sayres[127]; **Boston College** Erik P. Garrison[4]; **Broad Institute of MIT and Harvard** Robert E. Handsaker[13,40], Seva Kashin[13,40], Steven A. McCarroll[13,40]; **Cornell University:** Juan L. Rodriguez-Flores[45]; **European Molecular Biology Laboratory, European Bioinformatics Institute** Paul Flicek[12] (Principal Investigator), Laura Clarke[12], Xiangqun Zheng-Bradley[12]; **New York Genome Center** Yaniv Erlich[56,58], Melissa Gymrek[13,56,59,60], Thomas Frederick Willems[61]; **Stanford University** Carlos D. Bustamante[65] (Principal Investigator) (Co-Chair), Fernando L. Mendez[65], G. David Poznik[128], Peter A. Underhill[65]; **The Jackson Laboratory for Genomic Medicine** Charles Lee[18,19], Eliza Cerveira[18], Ankit Malhotra[18], Mallory Romanovitch[18], Chengsheng Zhang[18]; **University of Michigan** Gonçalo R. Abecasis[2] (Principal Investigator); **University of Queensland** Lachlan Coin[129] (Principal Investigator), Haojing Shao[129]; **Virginia Bioinformatics Institute** David Mittelman[130]; **Wellcome Trust Sanger Institute** Chris Tyler-Smith[4] (Principal Investigator) (Co-Chair), Qasim Ayub[4], Ruby Banerjee[4], Maria Cerezo[4], Yuan Chen[4], Thomas W. Fitzgerald[4], Sandra Louzada[4], Andrea Massaia[4], Shane McCarthy[4], Graham R. Ritchie[4], Yali Xue[4], Fengtang Yang[4]

Data coordination center group: Baylor College of Medicine Richard A. Gibbs[14] (Principal Investigator), Christie Kovar[14], Divya Kalra[14], Walker Hale[14], Donna Muzny[14], Jeffrey G. Reid[14]; **BGI-Shenzhen** Jun Wang[26,27,28,29,30] (Principal Investigator), Xu Dan[26], Xiaosen Guo[26,27], Guoqing Li[26], Yingrui Li[26], Chen Ye[26], Xiaole Zheng[26]; **Broad Institute of MIT and Harvard** David M. Altshuler[3]; **European Molecular Biology Laboratory, European Bioinformatics Institute** Paul Flicek[12] (Principal Investigator) (Co-Chair), Laura Clarke[12] (Project Lead), Xiangqun Zheng-Bradley[12]; **Illumina** David R. Bentley[5] (Principal Investigator), Anthony Cox[5], Sean Humphray[5], Scott Kahn[39]; **Max Planck Institute for Molecular Genetics** Ralf Sudbrak[32] (Project Lead), Marcus W. Albrecht[33], Matthias Lienhard[20]; **McDonnell Genome Institute at Washington University** David Larson[22]; **Translational Genomics Research Institute** David W. Craig[74] (Principal Investigator), Tyler Izatt[74], Ahmet A. Kurdoglu[74]; **US National Institutes of Health** Stephen T. Sherry[25] (Principal Investigator) (Co-Chair), Chunlin Xiao[25]; **University of California, Santa Cruz** David Haussler[83,84] (Principal Investigator); **University of Michigan** Gonçalo R. Abecasis[2] (Principal Investigator); **University of Oxford** Gil A. McVean[8,9] (Principal Investigator); **Wellcome Trust Sanger Institute** Richard M. Durbin[4] (Principal Investigator), Senduran Balasubramaniam[4], Thomas M. Keane[4], Shane McCarthy[4], James Stalker[4]

Samples and ELSI group: Aravinda Chakravarti[6] (Co-Chair), Bartha M. Knoppers[16] (Co-Chair), Gonçalo R. Abecasis[2], Kathleen C. Barnes[131], Christine Beiswanger[31], Esteban G. Burchard[80], Carlos D. Bustamante[65], Hongyu Cai[26], Hongzhi Cao[26,27], Richard M. Durbin[4], Norman P. Gerry[31], Neda Gharani[31], Richard A. Gibbs[14], Christopher R. Gignoux[80], Simon Gravel[54], Brenna Henn[132], Danielle Jones[44], Lynn Jorde[105], Jane S. Kaye[133], Alon Keinan[7], Alastair Kent[134], Angeliki Kerasidou[135], Yingrui Li[26], Rasika Mathias[136], Gil A. McVean[8,9], Andres Moreno-Estrada[65,69], Pilar N. Ossorio[137,138], Michael Parker[135], Alissa M. Resch[31], Charles N. Rotimi[139], Charmaine D. Royal[140], Karla Sandoval[65], Yeyang Su[26], Ralf Sudbrak[32], Zhongming Tian[26], Sarah Tishkoff[141], Lorraine H. Toji[31], Chris Tyler-Smith[4], Marc Via[142], Yuhong Wang[26], Huanming Yang[26], Ling Yang[26], Jiayong Zhu[26]

Sample collection: British from England and Scotland (GBR) Walter Bodmer[143]; **Colombians in Medellín, Colombia (CLM)** Gabriel Bedoya[144], Andres Ruiz-Linares[86]; **Han Chinese South (CHS)** Zhiming Cai[26], Yang Gao[145], Jiayou Chu[146]; **Finnish in Finland (FIN)** Leena Peltonen‡; **Iberian Populations in Spain (IBS)** Andres Garcia-Montero[147], Alberto Orfao[147]; **Puerto Ricans in Puerto Rico (PUR)** Julie Dutil[148], Juan C. Martinez-Cruzado[104], Taras K. Oleksyk[104]; **African Caribbean in Barbados (ACB)** Kathleen C. Barnes[131], Rasika A. Mathias[136], Anselm Hennis[149,150], Harold Watson[150], Colin McKenzie[151]; **Bengali in Bangladesh (BEB)** Firdausi Qadri[152], Regina LaRocque[152], Pardis C. Sabeti[13,48]; **Chinese Dai in Xishuangbanna, China (CDX)** Jiayong Zhu[26], Xiaoyan Deng[153]; **Esan in Nigeria (ESN)** Pardis C. Sabeti[13,48], Danny Asogun[154], Onikepe Folarin[155], Christian Happi[155,156], Omonwunmi Omoniwa[155,156], Matt Stremlau[13,48], Ridhi Tariyal[13,48]; **Gambian in Western Division – Mandinka (GWD)** Muminatou Jallow[8,157], Fatoumatta Sisay Joof[8,157], Tumani Corrah[8,157], Kirk Rockett[8,157], Dominic Kwiatkowski[8,157]; **Indian Telugu in the UK (ITU) and Sri Lankan Tamil in the UK (STU)** Jaspal Kooner[158]; **Kinh in Ho Chi Minh City, Vietnam (KHV)** Trần Tịnh Hiền[159], Sarah J. Dunstan[159,160], Nguyen Thuy Hang[159]; **Mende in Sierra Leone (MSL)** Richard Fonnie[161], Robert Garry[162], Lansana Kanneh[161], Lina Moses[162], Pardis C. Sabeti[13,48], John Schieffelin[162], Donald S. Grant[161,162]; **Peruvian in Lima, Peru (PEL)** Carla Gallo[163], Giovanni Poletti[163]; **Punjabi in Lahore, Pakistan (PJL)** Danish Saleheen[164,165], Asif Rasheed[164]

Scientific management: Lisa D. Brooks[166], Adam L. Felsenfeld[166], Jean E. McEwen[166], Yekaterina Vaydylevich[166], Eric D. Green[15], Audrey Duncanson[167], Michael Dunn[167], Jeffery A. Schloss[166], Jun Wang[26,27,28,29,30], Huanming Yang[26,168]

Writing group: Adam Auton[1], Lisa D. Brooks[166], Richard M. Durbin[4], Erik P. Garrison[4], Hyun Min Kang[2], Jan O. Korbel[12,17], Jonathan L. Marchini[8,9], Shane McCarthy[4], Gil A. McVean[8,9], Gonçalo R. Abecasis[2]

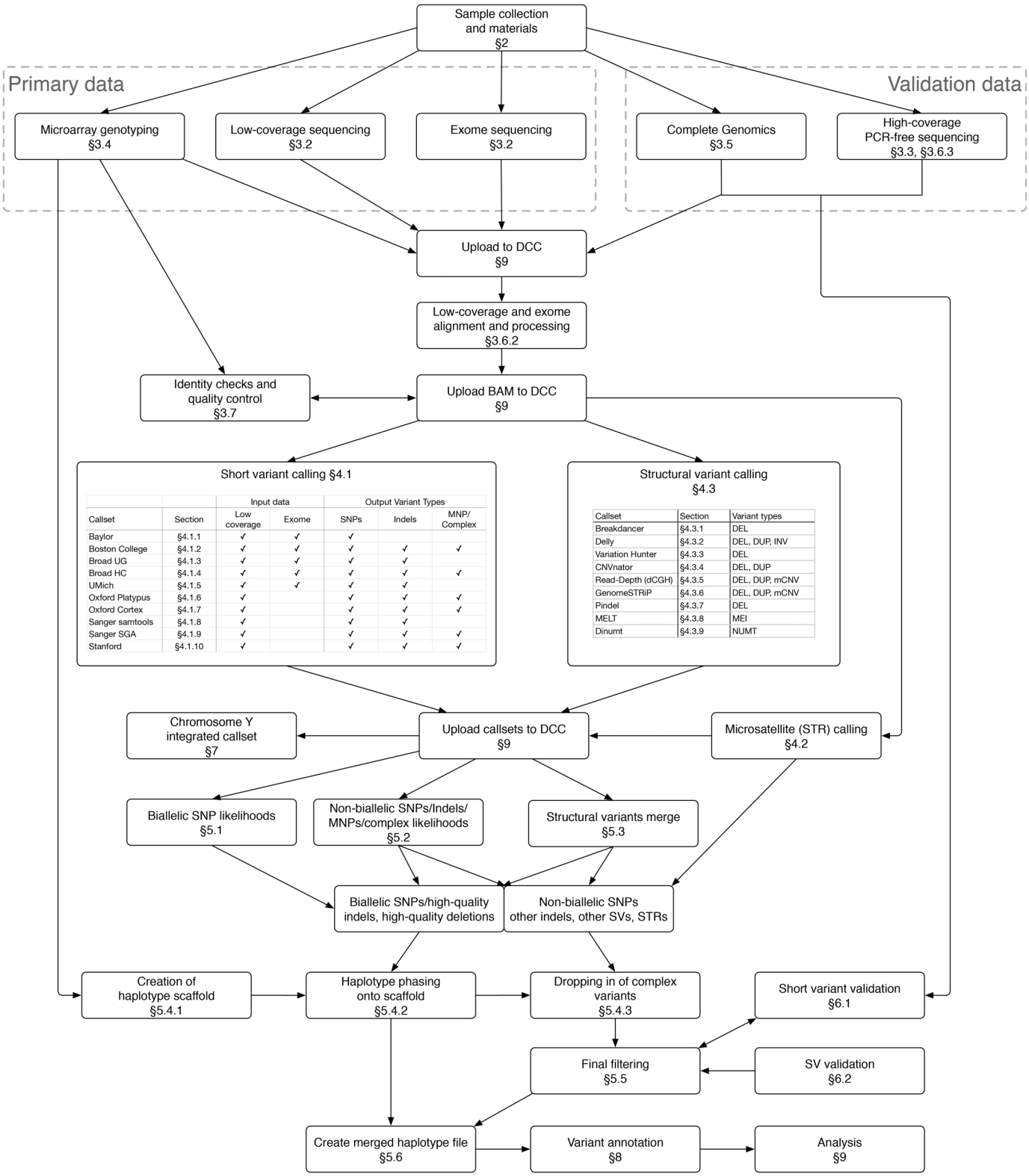
[1]Department of Genetics, Albert Einstein College of Medicine, Bronx, New York 10461, USA. [2]Center for Statistical Genetics, Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA. [3]Vertex Pharmaceuticals, Boston, Massachusetts 02210, USA. [4]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK. [5]Illumina United Kingdom, Chesterford Research Park, Little Chesterford, Nr Saffron Walden, Essex CB10 1XL, UK. [6]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. [7]Center for Comparative and Population Genomics, Cornell University, Ithaca, New York 14850, USA. [8]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. [9]Department of Statistics, University of Oxford, Oxford OX1 3TG, UK. [10]Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA. [11]Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA. [12]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. [13]The Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [14]Baylor College of Medicine, Human Genome Sequencing Center, Houston, Texas 77030, USA. [15]US National Institutes of Health, National Human Genome Research Institute, 31 Center Drive, Bethesda, Maryland 20892, USA. [16]Centre of Genomics and Policy, McGill University, Montreal, Quebec H3A 1A4, Canada. [17]European Molecular Biology Laboratory, Genome Biology Research Unit, Meyerhofstr. 1, Heidelberg, Germany. [18]The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, Connecticut 06032, USA. [19]Department of Life Sciences, Ewha Womans University, Ewhayeodae-gil, Seodaemun-gu, Seoul, South Korea 120-750. [20]Max Planck Institute for Molecular Genetics, D-14195 Berlin-Dahlem, Germany. [21]Dahlem Centre for Genome Research and Medical Systems Biology, D-14195 Berlin-Dahlem, Germany. [22]McDonnell Genome Institute at Washington University, Washington University School of Medicine, St Louis, Missouri 63108, USA. [23]USTAR Center for Genetic Discovery & Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, Utah 84112, USA. [24]Affymetrix, Santa Clara, California 95051, USA. [25]US National Institutes of Health, National Center for Biotechnology Information, 45 Center Drive, Bethesda, Maryland 20892, USA. [26]BGI-Shenzhen, Shenzhen 518083, China. [27]Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen, Denmark. [28]Princess Al Jawhara Albrahim Center of Excellence in the Research of Hereditary Disorders, King Abdulaziz University, Jeddah 80205, Saudi Arabia. [29]Macau University of Science and Technology, Avenida Wai long, Taipa, Macau 999078, China. [30]Department of Medicine and State Key Laboratory of Pharmaceutical Biotechnology, University of Hong Kong, 21 Sassoon Road, Hong Kong. [31]Coriell Institute for Medical Research, Camden, New Jersey 08103, USA. [32]European Centre for Public Heath Genomics, UNU-MERIT, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands. [33]Alacris Theranostics, D-14195 Berlin-Dahlem, Germany. [34]Personalis, Menlo Park, California 94025, USA. [35]US National Institutes of Health, National Human Genome Research Institute, 50 South Drive, Bethesda, Maryland 20892, USA. [36]Department of Computer Engineering, Bilkent University, TR-06800 Bilkent, Ankara, Turkey. [37]Seven Bridges Genomics, 1 Broadway, 14th floor, Cambridge, Massachusetts 02142, USA. [38]Department of Agronomy, Kansas State University, Manhattan, Kansas 66506, USA. [39]Illumina, San Diego, California 92122, USA.

©2015 Macmillan Publishers Limited. All rights reserved

[40]Department of Genetics, Harvard Medical School, Cambridge, Massachusetts 02142, USA. [41]SynapDx, Four Hartwell Place, Lexington, Massachusetts 02421, USA. [42]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA. [43]Seaver Autism Center and Department of Psychiatry, Mount Sinai School of Medicine, New York, New York 10029, USA. [44]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853, USA. [45]Department of Genetic Medicine, Weill Cornell Medical College, New York, New York 10044, USA. [46]European Molecular Biology Laboratory, Genomics Core Facility, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [47]Bill Lyons Informatics Centre, UCL Cancer Institute, University College London, London WC1E 6DD, UK. [48]Center for Systems Biology and Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [49]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [50]Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK. [51]Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1498, New York, New York 10029-6574, USA. [52]Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana 70803, USA. [53]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [54]McGill University and Genome Quebec Innovation Centre, 740, Avenue du Dr. Penfield, Montreal, Quebec H3A 0G1, Canada. [55]National Eye Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [56]New York Genome Center, 101 Avenue of the Americas, 7th floor, New York, New York 10013, USA. [57]Department of Systems Biology, Columbia University, New York, NY 10032, USA. [58]Department of Computer Science, Fu Foundation School of Engineering, Columbia University, New York, New York, USA. [59]Harvard–MIT Division of Health Sciences and Technology, Cambridge, Massachusetts 02139, USA. [60]General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA. [61]Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA.[62]Ontario Institute for Cancer Research, MaRS Centre, 661 University Avenue, Suite 510, Toronto, Ontario, M5G 0A3, Canada. [63]Department of Anthropology, Penn State University, University Park, Pennsylvania 16802, USA. [64]Rutgers Cancer Institute of New Jersey, New Brunswick, New Jersey 08903, USA. [65]Department of Genetics, Stanford University, Stanford, California 94305, USA. [66]Departments of Genetics and Pathology, Stanford University, Stanford, California 94305-5324, USA. [67]Ancestry.com, San Francisco, California 94107, USA. [68]DNAnexus, 1975 West El Camino Real STE 101, Mountain View California 94040, USA. [69]Laboratorio Nacional de Genómica para la Biodiversidad (LANGEBIO), CINVESTAV, Irapuato, Guanajuato 36821, Mexico. [70]Blavatnik School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel. [71]Department of Microbiology, Tel-Aviv University, Tel-Aviv 69978, Israel. [72]International Computer Science Institute, Berkeley, California 94704, USA. [73]Thermo Fisher Scientific, 200 Oyster Point Boulevard, South San Francisco, California 94080, USA. [74]The Translational Genomics Research Institute, Phoenix, Arizona 85004, USA. [75]Life Technologies, Beverly, Massachusetts 01915, USA. [76]Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, California 90024, USA. [77]Department of Psychiatry, University of California, San Diego, La Jolla, California 92093, USA. [78]Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, California 92093, USA. [79]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, New York 10461, USA. [80]Departments of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, California 94158, USA. [81]Institute for Quantitative Biosciences (QB3), University of California, San Francisco, 1700 4th Street, San Francisco, California 94158, USA. [82]Institute for Human Genetics, University of California, San Francisco, 1700 4th Street, San Francisco, California 94158, USA. [83]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, Santa Cruz, California 95064, USA. [84]Howard Hughes Medical Institute, Santa Cruz, California 95064, USA. [85]Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA. [86]Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK. [87]Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland. [88]Institute for Genetics and Genomics in Geneva, University of Geneva, 1211 Geneva, Switzerland. [89]Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland. [90]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA. [91]Department of Computational Medicine and Bioinfomatics, University of Michigan, Ann Arbor, Michigan 48109, USA. [92]Department of Human Genetics, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA. [93]Department of Pediatrics, University of Pittsburgh, Pittsburgh, Pennsylvania 15224, USA. [94]The University of Texas Health Science Center at Houston, Houston, Texas 77030, USA. [95]Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA. [96]University of Michigan Sequencing Core, University of Michigan, Ann Arbor, Michigan 48109, USA. [97]Istituto di Ricerca Genetica e Biomedica, CNR, Monserrato, 09042 Cagliari, Italy. [98]Dipartimento di Scienze Biomediche, Università delgi Studi di Sassari, 07100 Sassari, Italy. [99]University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, Texas 75390, USA. [100]Department of Pediatrics, University of Montreal, Ste. Justine Hospital Research Centre, Montreal, Quebec H3T 1C5, Canada. [101]Department of Genetics, Department of Biostatistics, Department of Computer Science, University of Chapel Hill, North Carolina 27599, USA. [102]Department of Bioinformatics and Genomics, College of Computing and Informatics, University of North Carolina at Charlotte, 9201 University City Boulevard, Charlotte, North Carolina 28223, USA. [103]Department of Medical Genetics, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, The Netherlands. [104]Department of Biology, University of Puerto Rico at Mayagüez, Mayagüez, Puerto Rico 00680, USA. [105]Eccles Institute of Human Genetics, University of Utah School of Medicine, Salt Lake City, Utah 84112, USA. [106]Department of Genetics, Rutgers University, Piscataway, New Jersey 08854, USA. [107]Department of Medicine, Division of Medical Genetics, University of Washington, Seattle, Washington 98195, USA. [108]Department of Biostatistics, University of Washington, Seattle, Washington 98195, USA. [109]Department of Physiology and Biophysics, Weill Cornell Medical College, New York, New York 10065, USA. [110]Department of Human Genetics, Radboud Institute for Molecular Life Sciences and Donders Centre for Neuroscience, Radboud University Medical Center, Geert Grooteplein 10, 6525 GA Nijmegen, The Netherlands. [111]Department of Molecular Developmental Biology, Faculty of Science, Radboud Institute for Molecular Life Sciences (RIMLS), Radboud University, 6500 HB Nijmegen, The Netherlands. [112]Institute of Genetics and Biophysics, National Research Council (CNR), 80125 Naples, Italy. [113]Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA. [114]Department of Computer Science, Yale University, New Haven, Connecticut 06520, USA. [115]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA. [116]Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota 55905, USA. [117]Department of Chemistry, Yale University, New Haven, Connecticut 06520, USA. [118]Molecular Epidemiology Section, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center 2333 ZA, The Netherlands. [119]Department of Computer Science, University of California, San Diego, La Jolla, California 92093, USA. [120]Beyster Center for Genomics of Psychiatric Diseases, University of California, San Diego, La Jolla, California 92093, USA. [121]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77230, USA. [122]Bina Technologies, Roche Sequencing, Redwood City, California 94065, USA. [123]Department of Surgery, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [124]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kemitorvet Building 208, 2800 Lyngby, Denmark. [125]Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, New York 10024, USA. [126]Department of Invertebrate Zoology, American Museum of Natural History, New York, New York 10024, USA. [127]School of Life Sciences, Arizona State University, Tempe, Arizona 85287-4701, USA. [128]Program in Biomedical Informatics, Stanford University, Stanford, California 94305, USA. [129]Institute for Molecular Bioscience, University of Queensland, St Lucia, QLD 4072, Australia. [130]Virginia Bioinformatics Institute, 1015 Life Sciences Drive, Blacksburg, Virginia 24061, USA. [131]Division of Allergy and Clinical Immunology, School of Medicine, Johns Hopkins University, Baltimore, Maryland 21205, USA. [132]Department of Ecology and Evolution, Stony Brook University, Stony Brook, New York 11794, USA. [133]Centre for Health, Law and Emerging Technologies, University of Oxford, Oxford OX3 7LF, UK. [134]Genetic Alliance, London N1 3QP, UK. [135]The Ethox Center, Nuffield Department of Population Health, University of Oxford, Old Road Campus, OX3 7LF, UK. [136]Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. [137]Department of Medical History and Bioethics, Morgridge Institute for Research, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA. [138]University of Wisconsin Law School, Madison, Wisconsin 53706, USA. [139]US National Institutes of Health, Center for Research on Genomics and Global Health, National Human Genome Research Institute, 12 South Drive, Bethesda, Maryland 20892, USA. [140]Department of African & African American Studies, Duke University, Durham, North Carolina 27708, USA. [141]Department of Genetics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA. [142]Department of Psychiatry and Clinical Psychobiology & Institute for Brain, Cognition and Behavior (IR3C), University of Barcelona, 08035 Barcelona, Spain. [143]Cancer and Immunogenetics Laboratory, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DS, UK. [144]Laboratory of Molecular Genetics, Institute of Biology, University of Antioquia, Medellín, Colombia. [145]Peking University Shenzhen Hospital, Shenzhen, 518036, China. [146]Institute of Medical Biology, Chinese Academy of Medical Sciences and Peking Union Medical College, Kunming 650118, China. [147]Instituto de Biologia Molecular y Celular del Cancer, Centro de Investigacion del Cancer/IBMCC (CSIC-USAL), Institute of Biomedical Research of Salamanca (IBSAL) & National DNA Bank Carlos III, University of Salamanca, 37007 Salamanca, Spain. [148]Ponce Research Institute, Ponce Health Sciences University, Ponce 00716, Puerto Rico. [149]Chronic Disease Research Centre, Tropical Medicine Research Institute, Cave Hill Campus, The University of the West Indies. [150]Faculty of Medical Sciences, Cave Hill Campus, The University of the West Indies. [151]Tropical Metabolism Research Unit, Tropical Medicine Research Institute, Mona Campus, The University of the West Indies. [152]International Centre for Diarrhoeal Disease Research, Dhaka, Bangladesh. [153]Xishuangbanna Health School, Xishuangbanna 666100, China. [154]Irrua Specialist Teaching Hospital, Edo State, Nigeria. [155]Redeemers University, Ogun State, Nigeria. [156]Harvard T. H. Chan School of Public Health, Boston, Massachusetts 02115, USA. [157]Medical Research Council Unit, The Gambia, Atlantic Boulevard, Fajara, P.O. Box 273, Banjul, The Gambia. [158]NHLI, Imperial College London, Hammersmith Hospital, London SW7 2AZ, UK. [159]Centre for Tropical Medicine, Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam. [160]Peter Doherty Institute of Infection and Immunity, The University of Melbourne, 792 Elizabeth Street, Melbourne VIC 3000, Australia. [161]Kenema Government Hospital, Ministry of Health and Sanitation, Kenema, Sierra Leone. [162]Tulane University Health Sciences Center, New Orleans, Louisiana 70112, USA. [163]Laboratorios de Investigación y Desarrollo, Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Peru. [164]Center for Non-Communicable Diseases, Karachi, Pakistan. [165]Department of Epidemiology and Biostatistics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [166]US National Institutes of Health, National Human Genome Research Institute, 5635 Fishers Lane, Bethesda, Maryland 20892, USA. [167]Wellcome Trust, Gibbs Building, 215 Euston Road, London NW1 2BE, UK. [168]James D. Watson Institute of Genome Sciences, Hangzhou 310008, China.

‡Deceased

©2015 Macmillan Publishers Limited. All rights reserved

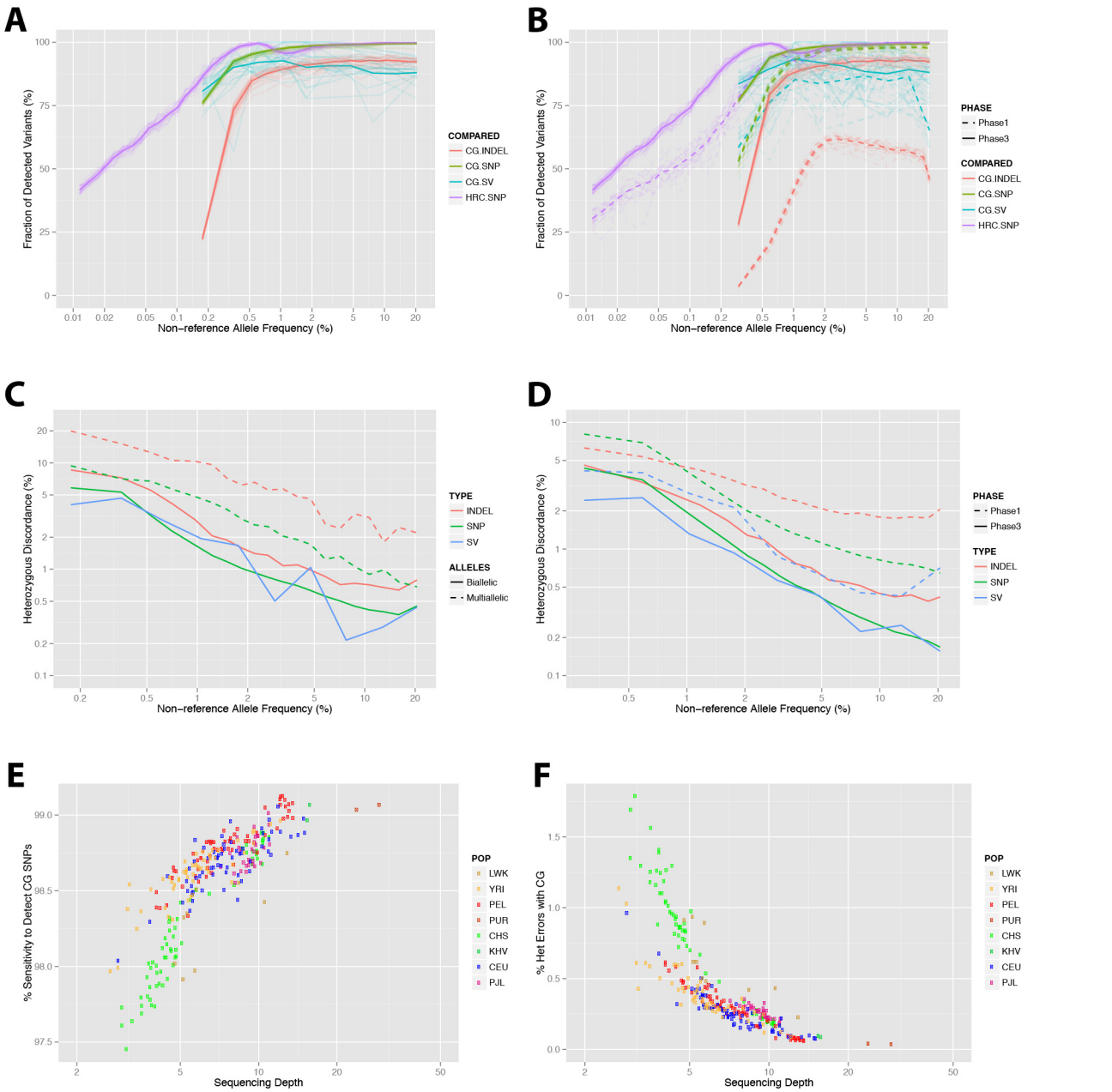


Callset	Section	Input data: Low coverage	Input data: Exome	Output Variant Types: SNPs	Output Variant Types: Indels	Output Variant Types: MNP/ Complex
Baylor	§4.1.1	✓	✓	✓		
Boston College	§4.1.2	✓	✓	✓	✓	✓
Broad UG	§4.1.3	✓	✓	✓	✓	
Broad HC	§4.1.4	✓	✓	✓	✓	✓
UMich	§4.1.5	✓	✓	✓	✓	
Oxford Platypus	§4.1.6	✓		✓	✓	✓
Oxford Cortex	§4.1.7	✓		✓	✓	✓
Sanger samtools	§4.1.8	✓		✓	✓	
Sanger SGA	§4.1.9	✓		✓	✓	✓
Stanford	§4.1.10	✓		✓	✓	✓

Callset	Section	Variant types
Breakdancer	§4.3.1	DEL
Delly	§4.3.2	DEL, DUP, INV
Variation Hunter	§4.3.3	DEL
CNVnator	§4.3.4	DEL, DUP
Read-Depth (dCGH)	§4.3.5	DEL, DUP, mCNV
GenomeSTRiP	§4.3.6	DEL, DUP, mCNV
Pindel	§4.3.7	DEL
MELT	§4.3.8	MEI
Dinumt	§4.3.9	NUMT

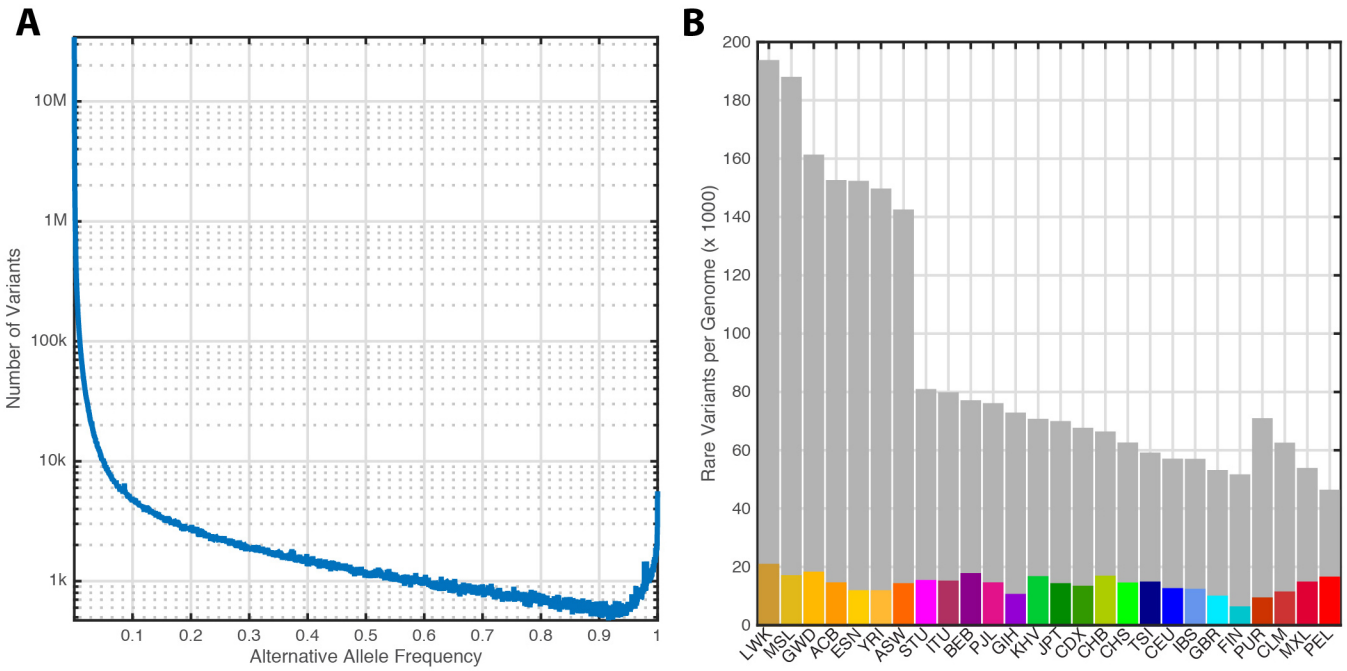
Extended Data Figure 1 | Summary of the callset generation pipeline. Boxes indicate steps in the process and numbers indicate the corresponding section(s) within the Supplementary Information.

©2015 Macmillan Publishers Limited. All rights reserved



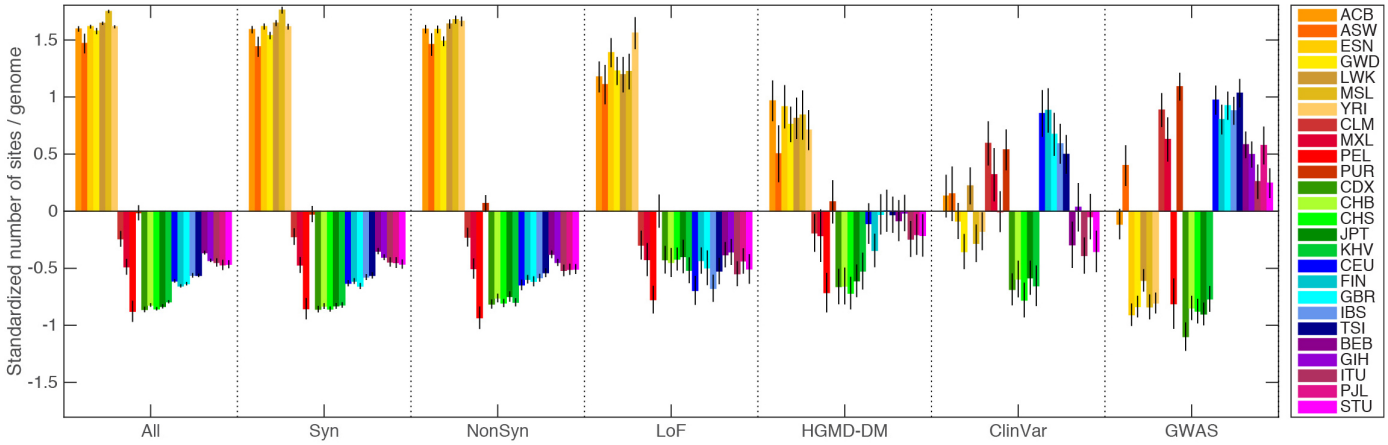
Extended Data Figure 2 | Power of discovery and heterozygote genotype discordance. **a**, The power of discovery within the main data set for SNPs and indels identified within an overlapping sample of 284 genomes sequenced to high coverage by Complete Genomics (CG), and against a panel of >60,000 haplotypes constructed by the Haplotype Reference Consortium (HRC)[9]. 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.

©2015 Macmillan Publishers Limited. All rights reserved



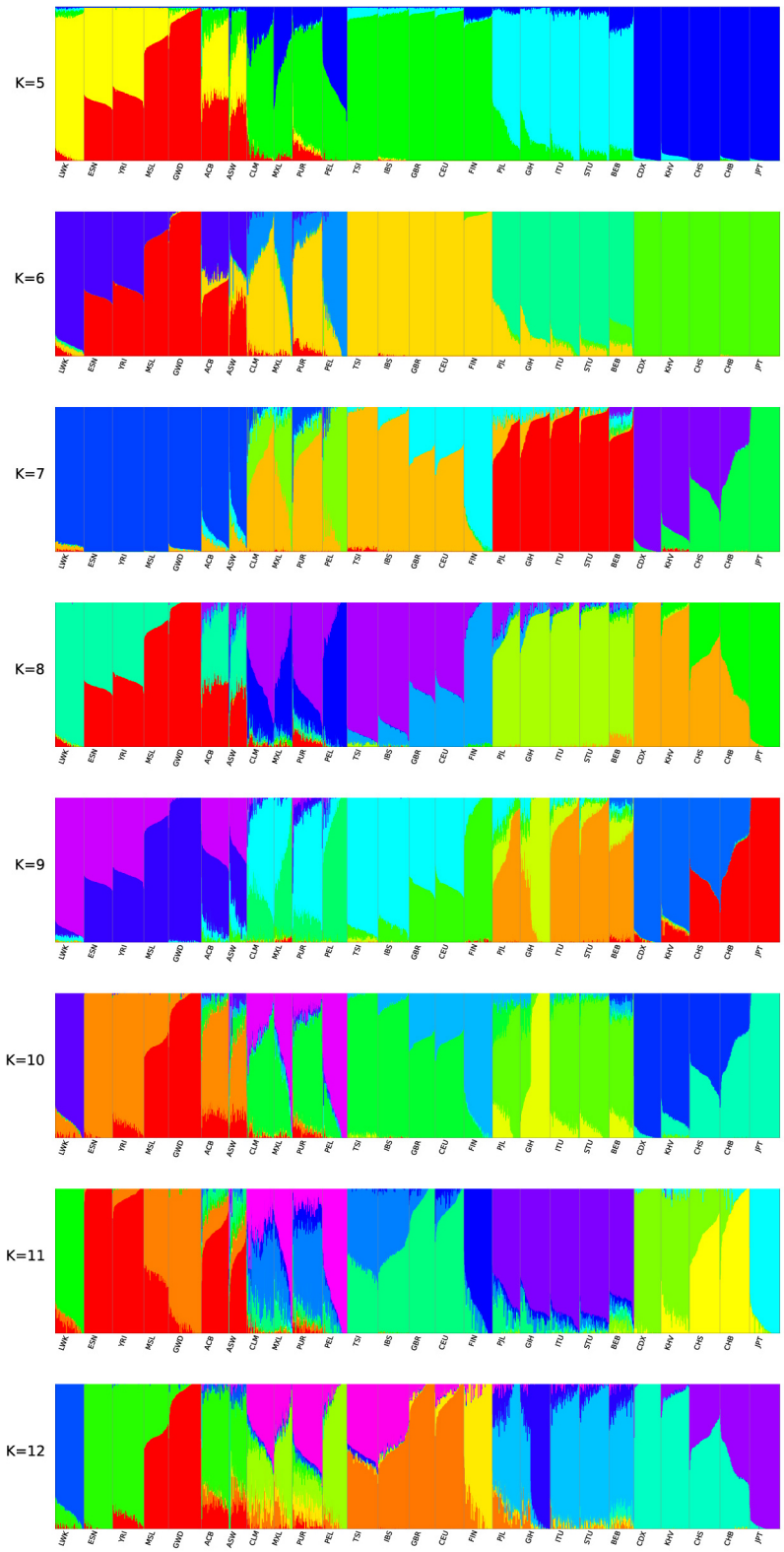
Extended Data Figure 3 | Variant counts. **a**, The number of variants within the phase 3 sample as a function of alternative allele frequency. **b**, The average number of detected variants per genome with whole-sample allele frequencies <0.5% (grey bars), with the average number of singletons indicated by colours.

©2015 Macmillan Publishers Limited. All rights reserved



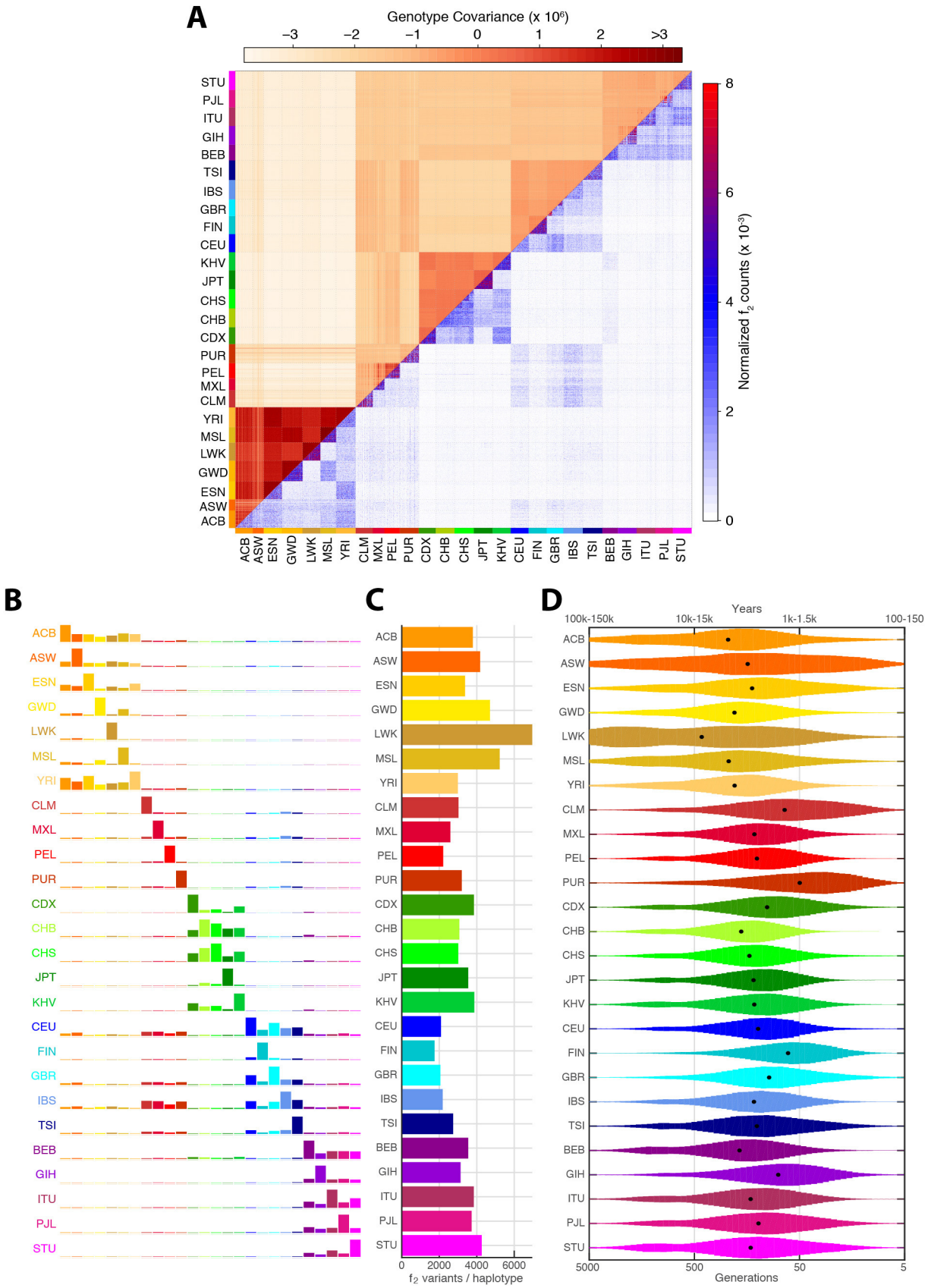
Extended Data Figure 4 | The standardized number of variant sites per genome, partitioned by population and variant category. For each category, *z*-scores were calculated by subtracting the mean number of sites per genome (calculated across the whole sample), and dividing by the standard deviation. 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.

©2015 Macmillan Publishers Limited. All rights reserved



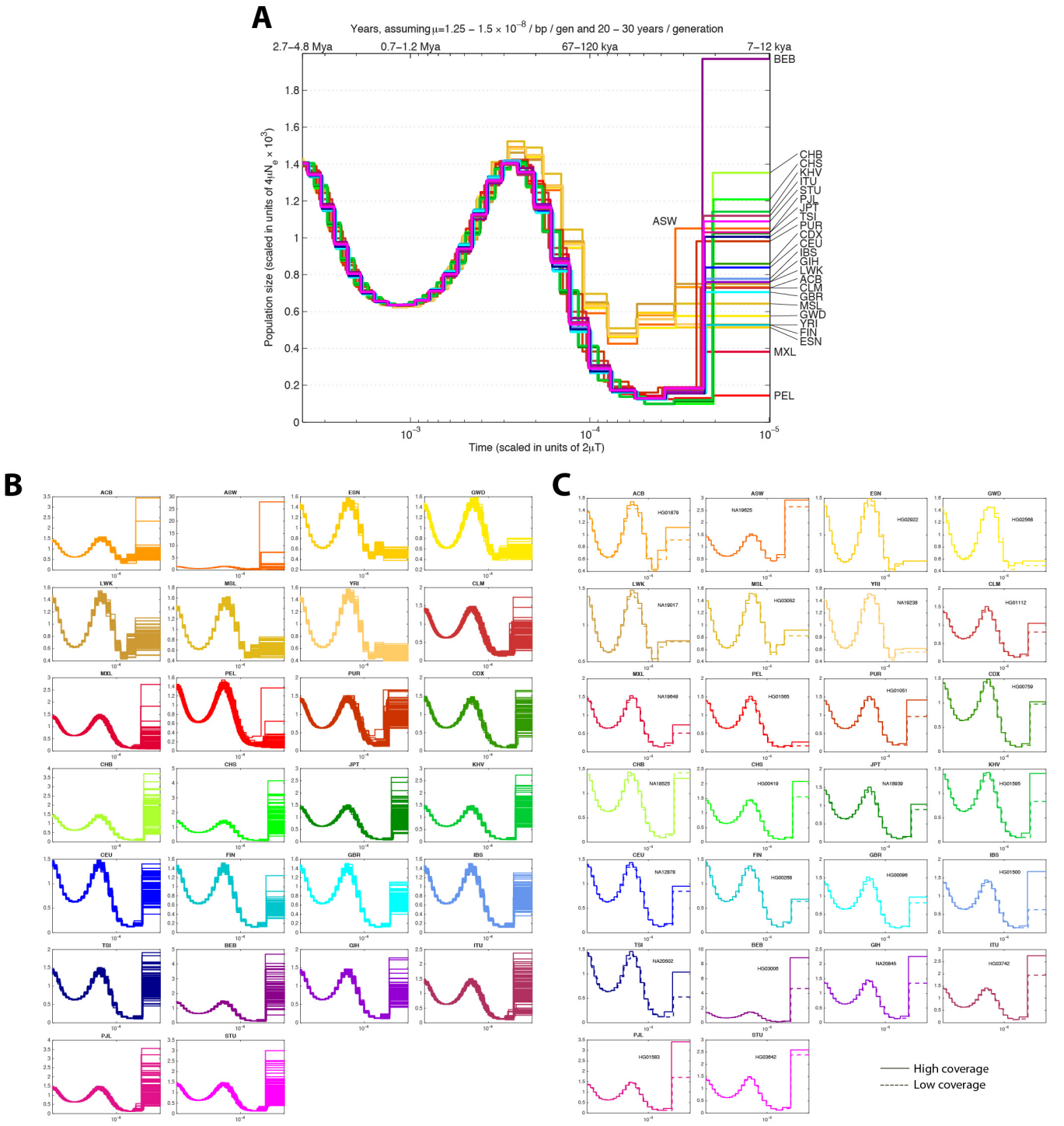
Extended Data Figure 5 | Population structure as inferred using the admixture program for $K = 5$ to 12.

©2015 Macmillan Publishers Limited. All rights reserved



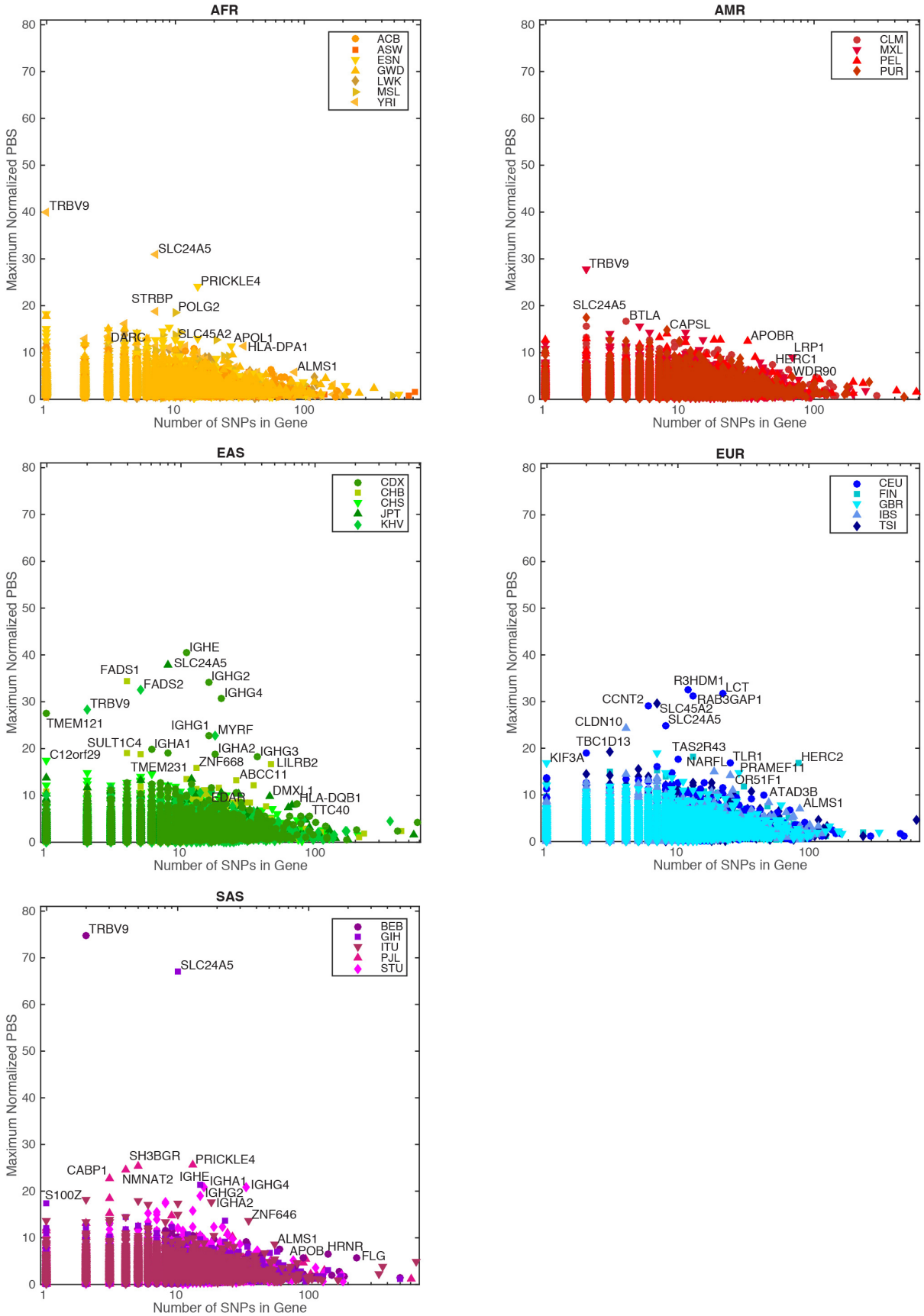
Extended Data Figure 6 | Allelic sharing. **a**, Genotype covariance (above diagonal) and sharing of f_2 variants (below diagonal) between pairs of individuals. **b**, Quantification of average f_2 sharing between populations. Each row represents the distribution of f_2 variants shared between individuals 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.

©2015 Macmillan Publishers Limited. All rights reserved



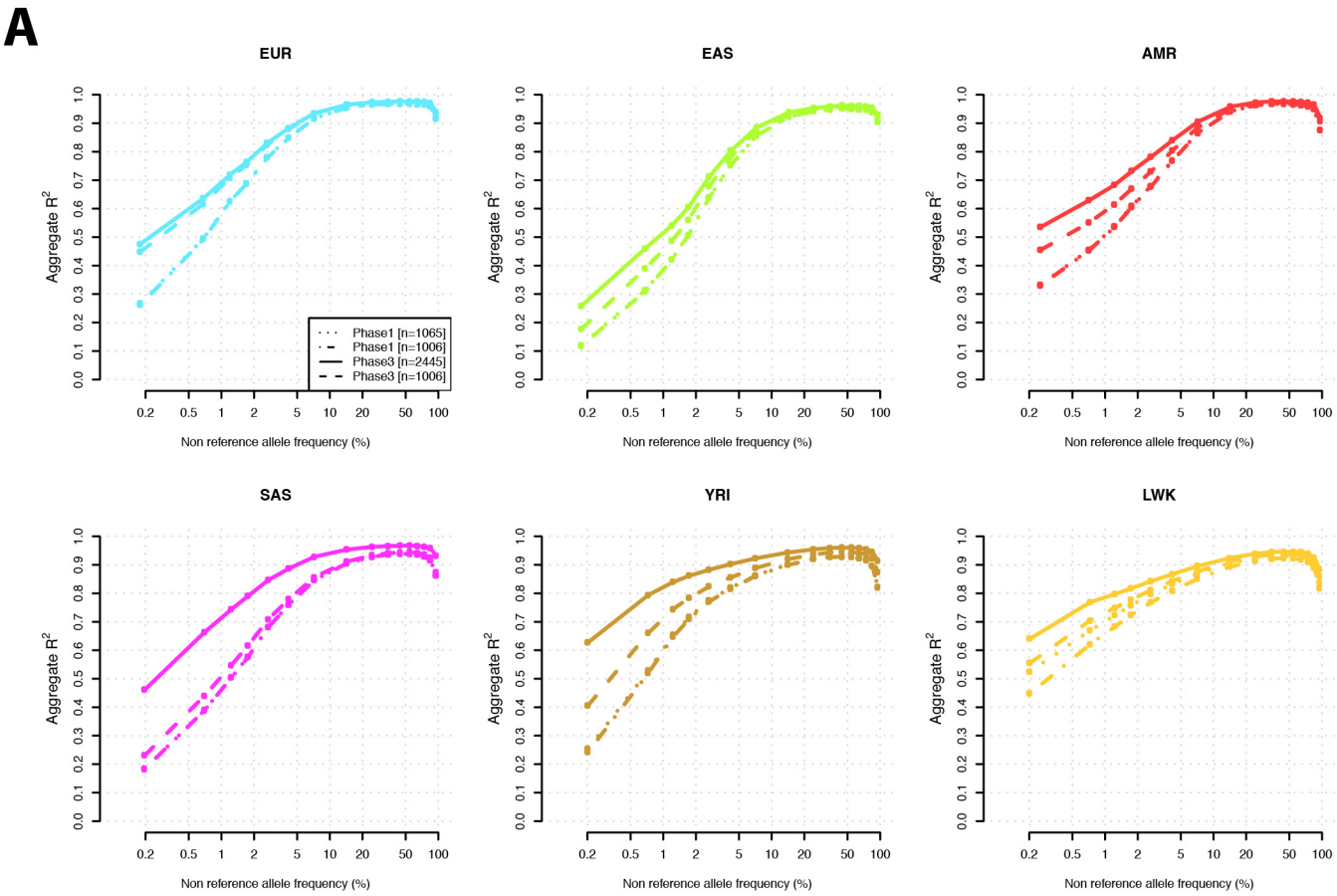
Extended Data Figure 7 | Unsmoothed PSMC curves. **a**, The median PSMC curve for each population. **b**, PSMC curves estimated separately for all individuals within the 1000 Genomes sample. **c**, Unsmoothed PSMC curves comparing estimates from the low coverage data (dashed lines) to those obtained 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.

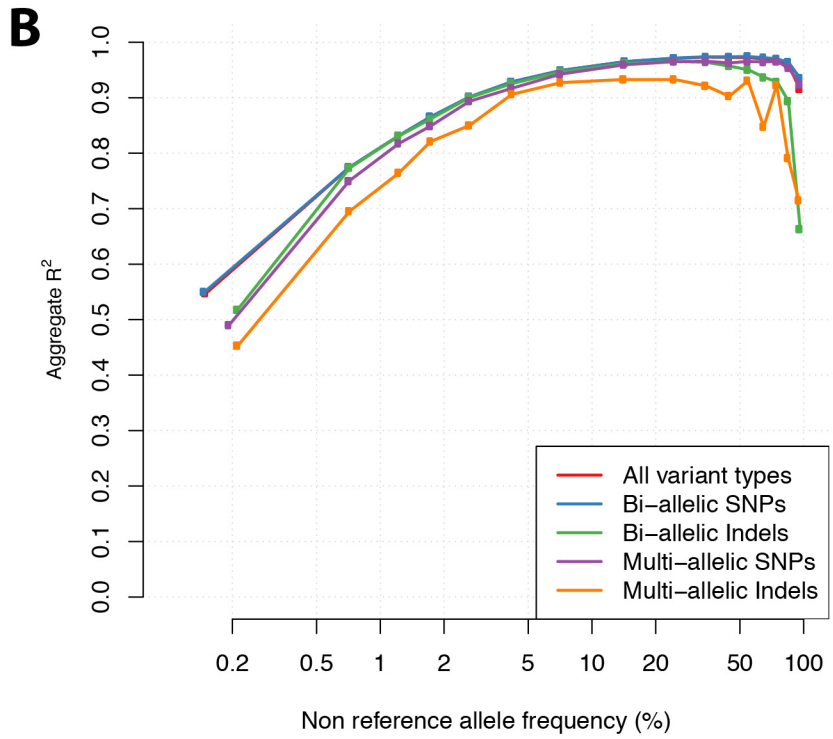
©2015 Macmillan Publishers Limited. All rights reserved



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.

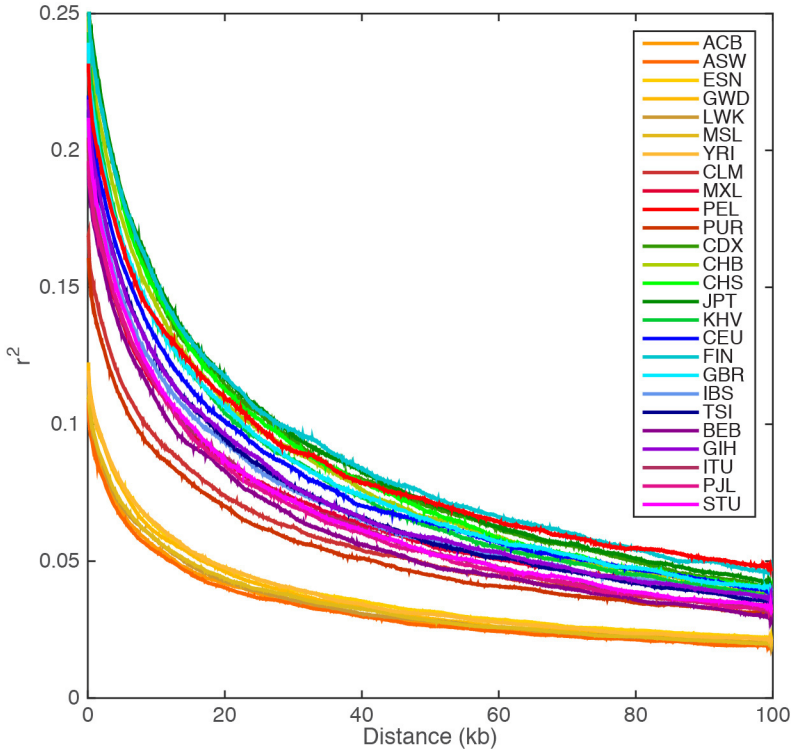
©2015 Macmillan Publishers Limited. All rights reserved





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.

©2015 Macmillan Publishers Limited. All rights reserved



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.

©2015 Macmillan Publishers Limited. All rights reserved