UKPMC Funders Group
Author Manuscript
Nature. Author manuscript; available in PMC 2011 April 1.

Published in final edited form as:
Nature. 2010 October 28; 467(7319): 1061–1073. doi:10.1038/nature09534.

A map of human genome variation from population scale sequencing

The 1000 Genomes Project Consortium

Abstract

The 1000 Genomes Project aims to provide a deep characterisation of human genome sequence variation as a foundation for investigating the relationship between genotype and phenotype. We present results of the pilot phase of the project, designed to develop and compare different strategies for genome wide sequencing with high throughput sequencing platforms. We undertook three projects: low coverage whole genome sequencing of 179 individuals from four populations, high coverage sequencing of two mother-father-child trios, and exon targeted sequencing of 697 individuals from seven populations. We describe the location, allele frequency and local haplotype structure of approximately 15 million SNPs, 1 million short insertions and deletions and 20,000 structural variants, the majority of which were previously undescribed. We show that over 95% of the currently accessible variants found in any individual are present in this dataset; on average, each person carries approximately 250 to 300 loss of function variants in annotated genes and 50 to 100 variants previously implicated in inherited disorders. We demonstrate how these results can be used to inform association and functional studies. From the two trios we directly estimate the rate of *de novo* germline base substitution mutations to be approximately 10^{-8} per base pair per generation. We find many putative functional variants with large allele frequency differences between populations. We explore the data with regard to signatures of natural selection, and identify a marked reduction of genetic variation in the neighbourhood of genes, due to selection at linked sites. These methods and public data will support the next phase of human genetic research.

Introduction

Understanding the relationship between genotype and phenotype is one of the central goals in biology and medicine. The reference human genome sequence[1] provides a foundation for the study of human genetics, but systematic investigation of human variation requires full knowledge of DNA sequence variation across the entire spectrum of allele frequencies and types of DNA differences. Substantial progress has already been made. By 2008 the public catalogue of variant sites (dbSNP 129) contained approximately 11 million single nucleotide polymorphisms (SNPs) and 3 million short insertions and deletions (indels)[2-4]. Databases of structural variants (SVs) (e.g., dbVAR) indexed the locations of large genomic variants. The International HapMap Project catalogued both allele frequencies and the correlation patterns between nearby variants, a phenomenon known as linkage disequilibrium (LD), across several populations for 3.5 million SNPs[3, 4].

These resources have driven disease gene discovery in the first generation of genome wide association studies (GWAS), wherein genotypes at several hundred thousand variant sites, combined with the knowledge of LD structure, allow the vast majority of common variants (here, those with > 5% minor allele frequency, or MAF) to be tested for association[4] with disease. Over the last five years association studies have identified more than a thousand genomic regions associated with disease susceptibility and other common traits[5]. Genome wide collections of both common and rare SVs have similarly been tested for association with disease[6].

UKPMC Funders Group Author Manuscript
UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

Despite these successes, much work is still needed to achieve a deep understanding of the genetic contribution to human phenotypes[7]. Once a region has been identified as harbouring a risk locus, detailed study of all genetic variants in the locus is required to discover the causal variant(s), to quantify their contribution to disease susceptibility, and to elucidate their roles in functional pathways. Low frequency and rare variants (here defined as 0.5% to 5% MAF, and below 0.5% MAF respectively) vastly outnumber common variants and also contribute significantly to the genetic architecture of disease but it has not yet been possible to study them systematically[7-9]. Meanwhile, advances in DNA sequencing technology have enabled the sequencing of individual genomes[10-13], illuminating the gaps in the first generation of databases that contain mostly common variant sites. A much more complete catalogue of human DNA variation is a prerequisite to fully understanding the role of common and low frequency variants in human phenotypic variation.

The aim of the 1000 Genomes Project is to discover, genotype and provide accurate haplotype information on all forms of human DNA polymorphism in multiple human populations. Specifically, the goal is to characterise over 95% of variants that are in genomic regions accessible to current high throughput sequencing technologies and that have allele frequency of 1% or higher (the classical definition of polymorphism) in each of five major population groups (populations in or with ancestry from Europe, East Asia, South Asia, West Africa and the Americas). Because functional alleles are often found in coding regions and have reduced allele frequencies, lower frequency alleles (down to 0.1%) will also be catalogued in such regions.

Here we report the results of the pilot phase of the project, the aim of which was to develop and compare different strategies for genome wide sequencing with high throughput platforms. To this end we undertook three projects: low coverage sequencing of 179 individuals, deep sequencing of six individuals in two trios, and exon sequencing of 1,000 genes in 697 individuals (Box 1). The results give us a much deeper, more uniform picture of human genetic variation than was previously available, enabling new insights into the landscapes of functional variation, genetic association and natural selection in humans.

Box 1

The 1000 Genomes pilot projects

To develop and assess multiple strategies to detect and genotype variants of various types and frequencies using high throughput sequencing we carried out three projects, using samples from the extended HapMap collection[14]

- **Trio** project: whole genome shotgun sequencing at high coverage (average 42x) of two families (one Yoruba from Ibadan, Nigeria (YRI), one of European ancestry in Utah (CEU)), each including two parents and one daughter. Each of the offspring was sequenced using three platforms and by multiple centres.
- **Low coverage** project: whole genome shotgun sequencing at low coverage (2-6x) of 59 unrelated individuals from YRI, 60 unrelated individuals from CEU, 30 unrelated Han Chinese individuals in Beijing (CHB) and 30 unrelated Japanese individuals in Tokyo (JPT).
- **Exon** project: targeted capture of the exons from nearly 1000 randomly selected genes (total of 1.4 Mb) followed by sequencing at high coverage (average > 50x) in 697 individuals from 7 populations of African (YRI, Luhya in Webuye, Kenya (LWK)), European (CEU, Toscani in Italia (TSI)) and East Asian (CHB, JPT, Chinese in Denver, Colorado (CHD)) ancestry.

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

The three experimental designs differ substantially both in their ability to obtain data for variants of different types and frequencies and in the analytical methods we used to infer individual genotypes. The Figure shows a schematic representation of the projects and the type of information obtained from each. Colours in the left region indicate different haplotypes in individual genomes, and line width indicates depth of coverage (not to scale). The shaded region to the right gives an example of genotype data that could be generated for the same sample under the three strategies (dots indicate missing data, dashes indicate phase information, i.e., whether heterozygous variants can be assigned to the correct haplotype). Within a short region of the genome, each individual carries two haplotypes, typically shared by others in the population. In the trio design, high sequence coverage and the use of multiple platforms enable accurate discovery of multiple variant types across most of the genome, with Mendelian transmission aiding genotype estimation, inference of haplotypes and quality control. The low coverage project, in contrast, efficiently identifies shared variants on common haplotypes[15, 16] (red or blue), but has lower power to detect rare haplotypes (light green) and associated variants (indicated by the missing alleles), and will give some inaccurate genotypes (indicated by the red allele incorrectly assigned G). The exon design enables accurate discovery of common, rare and low frequency variation in the targeted portion of the genome, but lacks the ability to observe variants outside the targeted regions or assign haplotype phase.

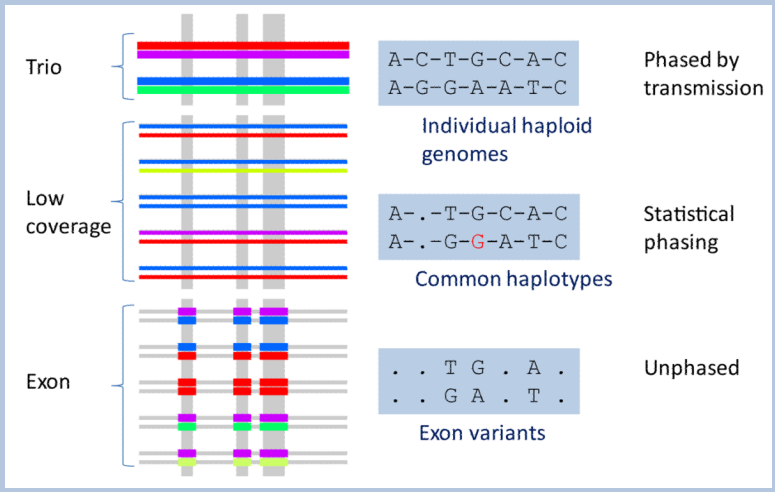

Figure.

Results

Overview of data generation, alignment and variant discovery

A total of 4.9 Tb of DNA sequence was generated in nine sequencing centres using three sequencing technologies, from DNA obtained from immortalised lymphoblastoid cell lines (Table 1 and Supplementary Table 1). All sequenced individuals provided informed consent and explicitly agreed to public dissemination of the variation data, as part of the HapMap Project (see Supplementary Information for details of informed consent and data release). The heterogeneity of the sequence data (read lengths from 25 to several hundred base pairs; single and paired end) reflects the diversity and rapid evolution of the underlying technologies during the project.

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

Analysis to detect and genotype sequence variants differed among variant types and the three projects, but all workflows shared four features:

- **Discovery**: alignment of sequence reads to the reference genome and identification of candidate sites or regions at which one or more samples differ from the reference sequence.
- **Filtering**: use of quality control measures to remove candidate sites that likely were false positives.
- **Genotyping**: estimation of the alleles present in each individual at variant sites or regions.
- **Validation**: assaying a subset of newly discovered variants using an independent technology, enabling the estimation of the false discovery rate. Independent data sources were used to estimate the accuracy of inferred genotypes.

All primary sequence reads, mapped reads, variant calls, inferred genotypes, estimated haplotypes and new independent validation data are publically available through the project website (www.1000genomes.org); filtered sets of variants, allele frequencies and genotypes were also deposited in dbSNP (www.ncbi.nlm.nih.gov/snp).

Alignment and the "accessible genome"—Sequencing reads were aligned to the NCBI36 reference genome (details in Supplementary Information) and made available in the BAM file format[17], an early innovation of the project for storing and sharing high throughput sequencing data. Accurate identification of genetic variation depends on alignment of the sequence data to the correct genomic location. We restricted most variant calling to the "accessible genome", defined as that portion of the reference sequence that remains after excluding regions with many ambiguously placed reads or unexpectedly high or low numbers of aligned reads (Supplementary Information). This approach balances the need to reduce incorrect alignments and false positive detection of variants against maximizing the proportion of the genome that can be interrogated

For the low coverage analysis, the accessible genome contains approximately 85% of the reference sequence and 93% of the coding sequences. Over 99% of sites genotyped in the second generation haplotype map (HapMap II)[4] are included. Of inaccessible sites, over 97% are annotated as high copy repeats or segmental duplications. However, only one quarter of previously discovered repeats and segmental duplications were inaccessible (Supplementary Table 2). Much of the data for the trio project was collected prior to technical improvements in our ability to map sequence reads robustly to some of the repeated regions of the genome (primarily longer, paired reads). For these reasons, stringent alignment was more difficult and a smaller portion of the genome was "accessible" in the trio project: 80% of the reference, 85% of coding sequence and 97% of HapMap II sites (Table 1).

Calibration, local realignment and assembly—The quality of variant calls is influenced by many factors including the quantification of base calling error rates in sequence reads, the accuracy of local read alignment and the method by which individual genotypes are defined. The project introduced key innovations in each of these areas (see Supplementary Information). First, base quality scores reported by the image processing software were empirically recalibrated by tallying the proportion that mismatched the reference sequence (at non-dbSNP sites) as a function of the reported quality score, position in read and other characteristics. Second, at potential variant sites local realignment of all reads was performed jointly across all samples, allowing for alternative alleles that contained indels. This realignment step substantially reduced errors, because local

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

misalignment, particularly around indels, can be a major source of error in variant calling. Finally, by initially analysing the data with multiple genotype and variant calling algorithms and then generating a consensus of these results, the project reduced genotyping error rates by 40-50% compared to those currently achievable using any one of the methods alone (Supplementary Figure 1).

We also used local realignment to generate candidate alternative haplotypes in the process of calling short (1-50 bp) indels[18], as well as local *de novo* assembly to resolve breakpoints for deletions greater than 50 bp. The latter resulted in a doubling of the number of large (> 1 kb) SVs delineated with base pair resolution[19]. Full genome *de novo* assembly was also performed (Supplementary Information), resulting in the identification of 3.7 Mb of novel sequence not matching the reference at a high threshold for assembly quality and novelty. All novel sequence matched other human and great ape sequences in the public databases.

Rates of variant discovery—In the trio project, with an average mapped sequence coverage of 42x per individual across six individuals and 2.3 Gb of accessible genome, we identified 5.9 million SNPs, 650,000 short indels (of 1-50 bp in length), and over 14,000 larger SVs. In the low coverage project, with average mapped coverage of 3.6x per individual across 179 individuals (Supplementary Fig. 2) and 2.4 Gb of accessible genome, we identified 14.4 million SNPs, 1.3 million short indels, and over 20,000 larger SVs. In the exon project, with an average mapped sequence coverage of 56x per individual across 697 individuals and a 1.4 Mb target, we identified 12,758 SNPs and 98 indels.

Experimental validation was used to estimate and control the false discovery rates (FDR) for novel variants (Supplementary Table 3). The FDR for each complete called set was controlled to be less than 5% for SNPs and short indels, and less than 10% for structural variants (FDR for novel variants 2.6% for trio SNPs, 10.9% for low coverage SNPs, 1.7% for low coverage indels; further data in Supplementary Information and Supplementary Tables 3, 4a and 4b).

Variation detected by the project is not evenly distributed across the genome: certain regions, such as the HLA and subtelomeric regions, show high rates of variation, while others, for example a 5 Mb gene dense and highly conserved region around 3p21, show very low levels of variation (Supplementary Fig. 3a). At the chromosomal scale we see strong correlation between different forms of variation, particularly between SNPs and indels (Supplementary Fig. 3b). However, we also find heterogeneity particular to types of SV, for example SVs resulting from nonallelic homologous recombination are apparently enriched in the HLA and in subtelomeric regions (Supplementary Fig. 3b, top).

Variant Novelty—As expected, the vast majority of sites variant in any given individual were already present in dbSNP; the proportion newly discovered differed substantially among populations, variant types and allele frequencies (Fig. 1). Novel SNPs had a strong tendency to be found only in one analysis panel (Fig. 1a). For SNPs also present in dbSNP version 129 (the last release prior to 1000 Genomes Project data), only 25% were specific to a single low coverage analysis panel and 56% were found in all panels. On the other hand, 84% of newly discovered SNPs were specific to a single analysis panel whereas only 4% were found in all analysis panels. In the exon project, where increased depth of coverage and sample size resulted in a higher fraction of low frequency variants among discovered sites, 96% of novel variants were restricted to samples from a single analysis panel. In contrast, many novel SVs were identified in all analysis panels, reflecting the lower degree of previous characterisation (Supplementary Figure 4).

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

Populations with African ancestry contributed the largest number of variants and contained the highest fraction of novel variants, reflecting the greater diversity in African populations. For example, 63% of novel SNPs in the low coverage project and 76% in the exon project were discovered in the African populations, compared to 20% and 33% in the European ancestry populations for the exon and low coverage projects respectively.

The larger sample sizes in the exon and low coverage projects allowed us to detect a large number of low frequency variants (MAF < 5%, Fig. 1b). Compared to the distribution expected from population genetic theory (the neutral coalescent with constant population size) we saw an excess of lower frequency variants in the exon project, reflecting purifying selection against weakly deleterious mutations and recent population growth. There are signs of a similar excess in the low coverage project SNPs, truncated below 5% variant allele frequency by reduction in power of our call set to discover variants in this range, as discussed further below.

As expected, nearly all of the high frequency SNPs discovered here were already present in dbSNP; this was particularly true in coding regions (Fig. 1c). The public databases were much less complete for SNPs at low frequencies, for short indels and for structural variants (Fig. 1d). For example, in contrast to coding SNPs (91% of common coding SNPs described here were already present in dbSNP), approximately 50% of common short indels observed in this project were novel. These results are expected given the sample sizes used in the sequencing efforts that discovered most of the SNPs previously in dbSNP, and the more limited, and lower resolution, efforts to characterize indels and larger structural variation across the genome.

The number of structural variants we observed declined rapidly with increasing variant length (Fig. 1d), with notable peaks corresponding to Alus and LINEs. The proportion of larger structural variants that was novel depended markedly on allele size, with variants 10 bp to 5 kb in size most likely to be novel (Fig. 1d). This is expected, as large (> 5 kb) deletions and duplications were previously discovered using array based approaches[14, 20], whereas smaller structural variants (apart from polymorphic Alu insertions) had been less well ascertained prior to this study.

Mitochondrial and Y chromosome sequences—Deep coverage of the mitochondrial genome allowed us to manually curate sequences for 163 samples (Supplementary Information). While variants that were fixed within an individual were consistent with the known phylogeny of the mitochondrial genome (Supplementary Fig. 5), we found a considerable amount of variation within individuals (heteroplasmy). For example, length heteroplasmy was detected in 79% of individuals compared with 52% using capillary sequencing[21], largely in the control region (Supplementary Fig. 6a). Base substitution heteroplasmy was observed in 45% of samples, seven times higher than reported in the control region alone[21], and was spread throughout the molecule (Supplementary Fig. 6b). The extent to which this heteroplasmy arose in cell culture remains unknown, but appears low (Supplementary Information).

The Y chromosome was sequenced at an average depth of 1.8x in the 77 males in the low coverage project, and 15.2x depth in the two trio fathers. Using customized analysis methods (Supplementary Information), we identified 2,870 variable sites, 74% novel, with 55/56 passing independent validation. The Y chromosome phylogeny derived from the new variants identified novel, well supported clades within some of the 12 major haplogroups represented among the samples (e.g., O2b in China and Japan; Supplementary Fig. 7). A striking pattern indicative of a recent rapid expansion specific to haplogroup R1b was observed, consistent with the postulated Neolithic origin of this haplogroup in Europe[22].

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

Power to detect variants

The ability of sequencing to detect a site that is segregating in the population is dominated by two factors: whether the nonreference allele is present among the individuals chosen for sequencing, and the number of high quality and well mapped reads that overlap the variant site in individuals who carry it. Simple models show that for a given total amount of sequencing, the number of variants discovered is maximised by sequencing many samples at low coverage[23, 24]. This is because high coverage of a few genomes, while providing the highest sensitivity and accuracy in genotyping a single individual, involves considerable redundancy and misses variation not represented by those samples. The low coverage project provides us with an empirical view of the power of low coverage sequencing to detect variants of different types and frequencies.

Fig. 2a shows the rate of discovery of variants in the CEU samples of the low coverage project as assessed by comparison to external data sources: HapMap and the exon project for SNPs and array CGH data[20] for large deletions. We estimate that while the low coverage project had only ~25% power to detect singleton SNPs, power to detect SNPs present five times in the 120 sampled chromosomes was ~90% (depending on the comparator), and power was essentially complete for those present 10 or more times. Similar results were seen in the YRI and CHB+JPT analysis panels at high allele counts, but slightly worse performance for variants present five times (~85% and 75% respectively at HapMap II sites; Supplementary Fig. 8). These results suggest that SNP discovery is less affected by the extent of LD (which is lowest in the YRI) than sequencing coverage (which was lowest in the CHB and JPT).

For deletions larger than 500 bp, power was approximately 40% for singletons and reached 90% for variants present ten times or more in the sample set. Our use of different algorithms for SV discovery ensured that all major mechanistic subclasses of deletions were found in our analyses (Supplementary Fig. 9). The lack of appropriate comparator datasets for short indels and larger structural variants other than deletions prevented a detailed assessment of the power to detect these types of variants. However, power to detect short indels was approximately 70% for variants present at least 5 times in the sample, based on the rediscovery of indels in samples overlapping with the SeattleSNPs project[25]. Extrapolating from comparisons to Alu insertions discovered in the Venter genome[26] suggested an average sensitivity for common mobile element insertions of about 75%. Analysis of a set of duplications[20] suggested only 30-40% of common duplications were discovered here, mostly as deletions with respect to the reference. Methods capable of discovering inversions and novel sequence insertions in low coverage data with comparable specificity remain to be developed.

In summary, low coverage shotgun sequencing provided modest power for singletons in each sample (~25-40%), and very good power for variants seen 5 or more times in the samples sequenced. We estimate that there was approximately 95% power to find SNPs with 5% allele frequency in the sequenced samples, and nearly 90% power to find SNPs with 5% allele frequency in populations related by 1% divergence (Fig. 2b). Thus we believe the projects found almost all accessible common variation in the sequenced populations and the vast majority of common variants in closely related populations.

Genotype accuracy

Genotypes, and, where possible, haplotypes, were inferred for all SNPs and short indels, and for most larger deletions (see Supplementary Information, and Table 1). For the low coverage data, statistically phased genotypes were derived by using LD structure in addition to sequence information at each site, in part guided by the HapMap 3 phased haplotypes.

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

SNP genotype accuracy varied considerably by pilot, coverage and allele frequency. In the low coverage project, the overall genotype error rate (based on a consensus of multiple methods) was 1-3% (Fig. 2c, Supplementary Fig. 10). The accuracy at heterozygous sites, a more sensitive measure than overall accuracy, was approximately 90% for the lowest frequency variants, increased to over 95% for intermediate frequencies and dropped to 70-80% for the highest frequency variants (i.e., those where the reference allele is the rare allele). We note that these numbers are derived from sites that can be genotyped using array technology, and performance may be lower in harder to access regions of the genome. We find only minor differences in genotype accuracy between populations, reflecting differences in coverage as well as haplotype diversity and extent of LD.

The accuracy of genotypes for large deletions was assessed against previous array based analyses[20] (Supplementary Fig. 11). The genotype error rate across all allele frequencies and genotypes was < 1%, with the accuracy of heterozygous genotypes at low (MAF < 3%), intermediate (MAF ~50%) and high frequency (MAF > 97%) variants estimated at 86%, 97% and 83% respectively. The greater apparent genotype accuracy of structural variants compared to SNPs in the low coverage project reflects the increased number of informative reads per individual for variants of large size and a bias in the known large deletion genotype set for larger, easier to genotype variants.

For calling genotypes in the low coverage samples, the utility of using LD information in addition to sequence data at each site was demonstrated by comparison to genotypes of the exon project, which were derived independently for each site using high coverage data. Fig. 2d shows the SNP genotype error rate as a function of depth at the genotyped sites in CEU. A similar number of variants was called, and at comparable accuracy, using minimum 4x depth in the low coverage project as was obtained with minimum 15x depth in the exon project. To genotype a high fraction of sites both projects needed to make calls at sites with low coverage, and the LD-based calling strategy for the low coverage project used imputation to make calls at nearly 15% more sites with only a modest increase in error rate.

The accuracy and completeness of the individual genome sequences in the low coverage project could be estimated from the trio mothers, each of whom was sequenced to high coverage, and for whom data subsampled to 4x were included in the low coverage analysis. Comparison of the SNP genotypes in the two projects showed that where the CEU mother had at least one variant allele according to the trio analysis, in 96.9% of cases the variant was also identified in the low coverage project and in 93.8% of cases the genotype was accurately inferred. For the YRI trio mother the equivalent figures are 95.0% and 88.4% respectively (note that false positives in the trio calls will lead to underestimates of the accuracy).

Putative functional variants

An individual's genome contains many variants of functional consequence, ranging from the beneficial to the highly deleterious. We estimated that an individual typically differs from the reference at 10,000-11,000 nonsynonymous sites (sequence differences that lead to differences in the protein sequence) in addition to 10,000-12,000 synonymous sites (differences in coding exons that do not lead to differences in the protein sequence; Table 2). We found a much smaller number of variants likely to have greater functional impact: in frame indels (190-210), premature stop codons (80-100), splice site disrupting variants (40-50), and deletions that shift reading frame (220-250), in each individual. We estimated that each genome is heterozygous for 50-100 variants classified by the Human Gene Mutation Database (HGMD) as causing inherited disorders (HGMD-DM). Estimates from the different pilot projects were consistent with each other, taking into consideration differences in power to detect low frequency variants, fraction of the accessible genome and

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

population differences (Table 2), as well as with previous observations based on personal genome sequences[10, 11]. Collectively, we refer to the 340-400 premature stops, splice site disruptions and frame shifts, affecting 250-300 genes per individual, as putative loss of function (LOF) variants.

In total, we found over 68,300 nonsynonymous SNPs, 34,161 of which were novel (Table 2). In an early analysis, 21,657 nonsynonymous SNPs were validated as polymorphic in 620 samples using a custom genotyping array (Table 2; Supplementary Information). The mean minor allele frequency in the array data was 2.2% for 4,573 novel variants, and 26.2% for previously discovered variants.

Overall we rediscovered 671 (1.3%) of the 50,361 coding single nucleotide variants in HGMD-DM (Supplementary Table 5). The types of disease for which variants were identified were biased towards certain categories (Supplementary Fig. 12), with diseases associated with the eye and reproduction significantly over represented and diseases of the nervous system significantly underrepresented. These biases reflect multiple factors including differences in the fitness effects of the variants, the extent of medical genetics research and differences in the false reporting rate among ‘disease causing’ variants.

As expected, and consistent with purifying selection, putative functional variants had an allele frequency spectrum depleted at higher allele frequencies, with putative LOF variants showing this effect more strongly (Supplementary Fig. 13). Of the low coverage nonsynonymous, stop-introducing, splice-disrupting and HGMD-DM variants, 67.3%, 77.3%, 82.2% and 84.7%, were private to single populations, compared to 61.1% for synonymous variants. Across these same functional classes, 15.8%, 25.9%, 21.6% and 19.9% of variants were found in only a single individual, compared to 11.8% of synonymous variants.

The tendency for deleterious functional variants to have lower allele frequencies has consequences for the discovery and analysis of this type of variation. In the deeply sequenced CEU trio father, who was not included in the low coverage project, 97.8% of all single base variants had been found in the low coverage project, but only 95% of nonsynonymous, 88% of stop inducing and 85% of HGMD-DM variants. The missed variants correspond to 389 nonsynonymous, 11 stop inducing and 13 HGMD-DM variants. As sample size increases, the number of novel variants per sequenced individual will decrease, but only slowly. Analyses based on the exon project data (Fig. 3) showed that on average 99% of the synonymous variants in an individual would be found in 100 deeply sequenced samples, whereas 250 samples would be required to find 99% of nonsynonymous variants and 320 samples would still find only 97.4% of the LOF variants present in an individual. Using detection power data from Fig. 2a, we estimated that 250 samples sequenced at low coverage would be needed to find 99% of the synonymous variants in an individual, and with 320 sequenced samples 98.5% of nonsynonymous and 96.3% of LOF variants would be found.

Application to association studies

Whole genome sequencing enables all genetic variants present in a sample set to be tested directly for association with a given disease or trait. To quantify the benefit of having more complete ascertainment of genetic variation beyond that achievable with genotyping arrays, we carried out expression quantitative trait loci (eQTL) association tests on the 142 low coverage samples for which expression data are available in the cell lines[27]. When association analysis (Spearman rank correlation, FDR < 5%, eQTLs within 50 kb of probe) was performed using all sites discovered in the low coverage project, a larger number of significant eQTLs (increase of ~20% to 50%) was observed as compared to association

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

analysis restricted to sites present on the Illumina 1M chip (Supplementary Table 6). The increase was lower in the CHB+JPT and CEU samples, where greater LD exists between previously examined and newly discovered variants, and higher in the YRI samples, where there are more novel variants and less LD. These results indicate that, while modern genotyping arrays capture most of the common variation, there remain substantial additional contributions to phenotypic variation from the variants not well captured by the arrays.

Population sequencing of large phenotyped cohorts will allow direct association tests for low frequency variants, with a resolution determined by the LD structure. An alternative that is less expensive, albeit less accurate, is to impute variants from a sequenced reference panel into previously genotyped samples[28, 29]. We evaluated the accuracy of imputation that used the current low coverage project haplotypes as the reference panel. Specifically, we compared genotypes derived by deep sequencing of one individual in each trio (the fathers) with genotypes derived using the HapMap 3 genotype data (which combined data from the Affymetrix 6.0 and Illumina 1M arrays) in those same two individuals and imputation based on the low coverage project haplotypes to fill in their missing genotypes. At variant sites (i.e., where the father was not homozygous for the reference), imputation accuracy was highest for SNPs at which the minor allele was observed at least 6 times in our low coverage samples, with an error rate of ~4% in CEU and ~10% in YRI, and became progressively worse for rarer SNPs, with error rates of 35% for sites where the minor allele was observed only twice in the low coverage samples (Fig. 4a).

Although the ability to impute rare variants accurately from the 1000 Genomes resource is currently limited, the completeness of the resource nevertheless increases power to detect association signals. To demonstrate the utility of imputation in disease samples, we imputed into an eQTL study of ~400 children of European ancestry[30] using the low coverage pilot data and HapMap II as reference panels. By comparison to directly genotyped sites we estimated that the effective sample size at variants imputed from the pilot CEU low coverage data set is 91% of the true sample size for variants with allele frequencies above 10%, 76% in the allele frequency range 4-6%, and 54% in the range 1-2%. Imputing over 6 million variants from the low coverage project data increased the number of detected cis-eQTLs by ~16%, compared to a 9% increase with imputing from HapMap II (FDR 5%, signal within 50 kb of transcript; for an example see Fig. 4b).

In addition to this modest increase in the number of discoveries, testing almost all common variants allows identification of many additional candidate variants that might underlie each association. For example, we find that rs11078928, a variant in a splice site for GSDMB, is in strong LD with SNPs near ORDML3 previously associated with asthma, Crohn's Disease, Type 1 Diabetes and rheumatoid arthritis, thus suggesting the hypothesis that GSDMB could be the causative gene in these associations. Although rs11078928 is not newly discovered, it was not included in HapMap or on commercial SNP arrays, and thus could not have been identified as associated with these diseases prior to this project. Similarly, a recent study[31] used project data to show that coding variants in *APOL1* likely underlie a major risk for kidney disease in African Americans previously attributed (at a lower effect size) to *MYH9*. These examples demonstrate the value of having much more complete information on LD, the almost complete set of variants in the regions, and putative functional variants in known association intervals.

Testing almost all common variants also allows us to examine general properties of genetic association signals. The NHGRI GWAS catalogue (www.genome.gov/gwastudies, accessed July 15, 2010) described 1,227 unique SNPs associated with one or more traits ($p < 5\times10^{-8}$). Of these, 1,185 (96.5%) are present in the low coverage CEU dataset. Under 30% of these are either annotated as nonsynonymous variants (77, 6.5%) or in substantial LD ($r^2 > 0.5$)

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

with a nonsynonymous variant (272, 23%). In the latter group, only 93 (8.4%) are in strong LD ($r^2 > 0.9$) with a nonsynonymous variant. Since we tested ~95% of common variation, these results suggest that no more than a third of complex trait association signals are likely to be caused by common coding variation. Although it remains to be seen whether reported associations are better explained through weak LD to coding variants with strong effects, these results are consistent with the view that most contributions of common variation to complex traits are regulatory in nature.

Mutation, recombination and natural selection

Project sequence data allowed us to investigate fundamental processes that shape human genetic variation including mutation, recombination and natural selection.

Detecting de novo mutations in trio samples—Deep sequencing of individuals within a pedigree offers the potential to detect *de novo* germline mutation events. Our approach was to allow a relatively high false discovery rate in an initial screen to capture a large fraction of true events, then use a second technology to rule out false positive mutations.

In the CEU and YRI trios respectively, 3,236 and 2,750 candidate *de novo* germline single base mutations were selected for further study, based on their presence in the child but not the parents. Of these, 1,001 (CEU) and 669 (YRI) were validated by resequencing the cell line DNA. When these were tested for segregation to offspring (CEU) or in non-clonal DNA from whole blood (YRI), only 49 CEU and 36 YRI candidates were confirmed as true germline mutations. Correcting for the fraction of the genome accessible to this analysis provided an estimate of the per generation base pair mutation rate of 1.2×10^{-8} and 1.0×10^{-8} in the CEU and YRI trios respectively. These values are similar to estimates obtained from indirect evolutionary comparisons[32], direct studies based on pathogenic mutations[33], and a recent analysis of a single family[34].

We infer that the remaining vast majority (952 CEU and 633 YRI) of the validated variants were somatic or cell line mutations. The greater number of these validated non-germline mutations in the CEU cell line perhaps reflects the greater age of the CEU cell culture. Across the two trio offspring, we observed a single, synonymous, coding germline mutation, and 17 coding non-germline mutations of which 16 were nonsynonymous, perhaps suggesting selection during cell culture.

Although the number of non-germline variants found per individual is a very small fraction of the total number of variants per individual (~0.03% for the CEU child and ~0.02% for the YRI child), these variants will not be shared between samples. Assuming that the number of non-germline mutations in these two trios is representative of all cell line DNA we analysed, we estimate that non-germline mutations might constitute 0.36% and 2.4% of all variants, and 0.61% and 3.1% of functional variants, in the low coverage and exon pilots respectively. In larger samples of thousands the overall false positive rates from cell line mutations would become significant, and confound interpretation, suggesting that large scale studies should use DNA from primary tissue such as blood where possible.

Constraint around genes and the effects of selection on local variation—Natural selection can affect levels of DNA variation across genes in multiple ways: strongly deleterious mutations will be rapidly eliminated by natural selection, weakly deleterious mutations can segregate in populations but rarely become fixed, and selection at nearby sites (both purifying and adaptive) can reduce genetic variation through background selection[35] and the hitchhiking effect[36]. The effect of these different forces on genetic variation can be disentangled by examining patterns of diversity and divergence within and around known

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

functional elements. The low coverage data enables, for the first time, genome wide analysis of such patterns in multiple populations. Fig. 5a (top) shows the pattern of diversity relative to genic regions measured by aggregating estimates of heterozygosity around protein coding genes. Within genes, exons harbor the least diversity (about 50% of that of introns) and 5′ and 3′ UTRs harbor slightly less diversity than immediate flanking regions and introns. However, this variation in diversity is fully explained by the level of divergence (Fig. 5a lower) consistent with the common part of the allele frequency spectrum being dominated by effectively neutral variants, and weakly deleterious variants contributing only to the rare end of the frequency spectrum.

In contrast, diversity in the immediate vicinity of genes (scaled by divergence) is reduced by approximately 10% relative to sites distant from any gene (Fig 5b). Although a similar reduction has been seen previously in gene dense regions[37], project data enable the scale of the effect to be determined. We find that the reduction extends up to 0.1cM away from genes, typically 120 kb, suggesting that selection at linked sites restricts variation relative to neutral levels across the majority of the human genome.

Positive selection and the distribution of genetic variation among populations —Previous inferences about demographic history and the role of local adaptation in shaping human genetic variation made from genome wide genotype data[4, 38, 39] have been limited by the partial and complex ascertainment of SNPs on genotype arrays. While data from the 1000 Genome Project pilots are neither fully comprehensive nor fully free of ascertainment bias (issues include low power for rare variants, noise in allele frequency estimates, some false positives, non-random data collection across samples, platforms and populations, and the use of imputed genotypes), they can be used to address key questions about the extent of differentiation among populations, the presence of highly differentiated variants and the ability to fine map signals of local adaptation.

Although the average level of population differentiation is low (at sites genotyped in all populations the mean value of Wright's Fst is 0.071 between CEU and YRI, 0.083 between YRI and CHB+JPT and 0.052 between CHB+JPT and CEU), we find several hundred thousand SNPs with large allele frequency differences in each population comparison (Fig. 5c). As seen in previous studies[4, 39], the most highly differentiated sites were enriched for nonsynonymous variants, suggestive of the action of local adaptation. The completeness of common variants in the low coverage resource enables new perspectives in the search for local adaptation. First, it provides a more comprehensive catalogue of fixed differences between populations, of which there are very few: two between CEU and CHB+JPT (including the A111T missense variant in SLC24A5[40] contributing to light skin colour), four between CEU and YRI (including the −46 GATA box null mutation upstream of *DARC*[41], the Duffy O allele leading to *vivax* malaria resistance) and 72 between CHB+JPT and YRI, including 24 around the exocyst complex component gene *EXOC6B*; see Supplementary Table 7 for a complete list. Second, it provides new candidates for selected variants, genes and pathways. For example, we identified 139 nonsynonymous (NS) variants showing large allele frequency differences (at least 0.8) between populations (Supplementary Table 8), including at least two genes involved in meiotic recombination, *FANCA* (9th most extreme NS SNP in CEU vs CHB+JPT) and *TEX15* (13th most extreme NS SNP in CEU vs YRI, and 26th most extreme NS SNP in CHB+JPT vs YRI). Because we are finding almost all common variants in each population, these lists should contain the vast majority of the near fixed differences among these populations. Finally, it enables fine mapping of the signals of local selective sweeps (Supplementary Fig. 14) and a characterisation of the footprint of such events on local variation. For example, we find that the signal of population differentiation around high differentiation genic SNPs is typically less than 0.2 cM (Fig. 5d),

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

suggesting that the signal of local adaptation should typically be resolved to one or a few genes.

The effect of recombination on local sequence evolution—We estimated a fine-scale genetic map from the phased low coverage genotypes. Recombination hotspots were narrower than previously estimated[4] (mean hotspot width of 2.3 kb compared to 5.5 kb in HapMap II; Fig. 6a), although, unexpectedly, the estimated average peak recombination rate in hotspots is lower in YRI (13 cM/Mb) than in CEU and CHB+JPT (20cM/Mb). In addition, crossover activity is less concentrated in the genome in YRI, with 70% of recombination occurring in 10% of the sequence rather than 80% of the recombination for CEU and CHB+JPT (Fig. 6b). A possible biological basis for these differences is that PRDM9, which binds a DNA motif strongly enriched in hotspots and influences the activity of LD-defined hotspots[42-45] , shows length variation in its DNA binding zinc fingers within populations, and substantial differentiation between African and non-African populations, with a greater allelic diversity in Africa[45]. This could mean greater diversity of hotspot locations within Africa and therefore a less concentrated picture in this data set of recombination and lower usage of LD-defined hotspots (which require evidence in at least two populations and therefore will not reflect hotspots present only in Africa).

The low coverage data also allowed us to address a longstanding debate about whether recombination has any local mutagenic effect. Direct examination of diversity around hotspots defined from LD data is potentially biased (because the detection of hotspots requires variation to be present), but we can without bias examine rates of SNP variation and recombination around the PRDM9 binding motif associated with hotspots. Fig. 6c shows the local recombination rate and pattern of SNP variation around the motif compared to the same plots around a motif that is a single base difference away. While the motif is associated with a sharp peak in recombination rate, there is no systematic effect on local rates of SNP variation. We infer that, although recombination may influence the fate of new mutations, for example through biased gene conversion, there is no evidence that it influences the rate at which new variants appear.

Discussion

The 1000 Genomes Project launched in 2008 with the goal of creating a public reference database for DNA polymorphism that is 95% complete at allele frequency 1%, and more complete for common variants and exonic variants, in each of multiple human population groups. The three pilot projects described here were designed to develop and evaluate methods to use high throughput sequencing to achieve these goals. The results indicate (a) that robust protocols now exist for generating both whole genome shotgun and targeted sequence data, (b) that algorithms to detect variants from each of these designs have been validated and (c) that low coverage sequencing offers an efficient approach to detect variation genome wide, whereas targeted sequencing offers an efficient approach to detect and accurately genotype rare variants in regions of functional interest (such as exons).

Data from the pilot projects are already informing medical genetic studies. As shown in our analysis of prior eQTL datasets, a more complete catalogue of genetic variation can identify signals previously missed and dramatically increase the number of identified candidate functional alleles at each locus. Project data have been used to impute over 6 million genetic variants into GWAS, for traits as diverse as smoking[46] and multiple sclerosis[47], as an exclusionary filter in Mendelian disease studies[48] and tumor sequencing studies, and to design the next generation of genotyping arrays.

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

The results from this study also provide a template for future genome-wide sequencing studies on larger sample sets. Our plans for achieving the 1000 Genomes Project goals are described in Box 2. Other studies using phenotyped samples are already using components of the design and analysis framework described above.

Measurement of human DNA variation is an essential prerequisite for carrying out human genetics research. The 1000 Genomes Project represents a step towards a complete description of polymorphic human DNA sequence variation. The larger dataset provided by the full 1000 Genomes Project will allow more accurate imputation of variants in GWAS and thus better localization of disease associated variants. The project will provide a template for studies using genome wide sequence data. Applications of these data, and the methods developed to generate them, will contribute to a much more comprehensive understanding of the role of inherited DNA variation in human history, evolution and disease.

Box 2

Design of the Full 1000 Genomes Project

The production phase of the full 1000 Genomes Project will combine low coverage whole genome sequencing, array based genotyping, and deep targeted sequencing of all coding regions in 2,500 individuals from five large regions of the world (five population samples of 100 in or with ancestry from each of Europe, East Asia, South Asia and West Africa, and seven populations totalling 500 from the Americas; Supplementary Table 9). We will increase the low coverage average depth to over 4x per individual, and use blood derived DNA where possible to minimise somatic and cell line false positives.

A clustered sampling approach was chosen to improve low frequency variant detection in comparison to a design in which a smaller number of populations were sampled to a greater depth. In a region containing a cluster of related populations, genetic drift can lead variants that are at low frequency overall to be more common (hence easily detectable) in one population but less common (hence likely to be undetectable) in another. We modelled this process using project data (see Supplementary Information) assuming that five sampled populations are equally closely related to each other (Fst = 1%). We found that the low coverage sequencing in this design would discover 95% of variants in the accessible genome at 1% frequency across each broad geographic region, between 90% and 95% of variants at 1% frequency in any one of the sampled populations and about 85% of variants at 1% frequency in any equally related but unsampled population. The chart shows predicted discovery curves for variants at different frequencies with details as for Fig. 2b. The model is conservative, in that it ignores migration and the contribution to discovery from more distantly related populations, each of which will increase sensitivity for variants in any given population. In exons, the full project should have 95% power to detect variants at a frequency of 0.3% and approximately 60% power for variants at a frequency of 0.1%.

In addition to improved detection power, we expect the full project to have increased genotype accuracy due to (a) advances in sequencing technology that are reducing per base error rates and alignment artefacts, (b) increased sample size, which improves imputation based methods, (c) ongoing algorithmic improvements, and (d) the designing by the project of genotyping assays that will directly genotype up to 10 million common and low frequency variants (SNPs, indels and SVs) observed in the low coverage data. In addition, we expect the fraction of the genome that is accessible to increase. Longer read lengths, improved protocols for generating paired reads, and the use of more powerful assembly based alignment methods are expected to increase accessibility from 80-85% to above 90% of the reference genome (Supplementary Fig. 15).

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

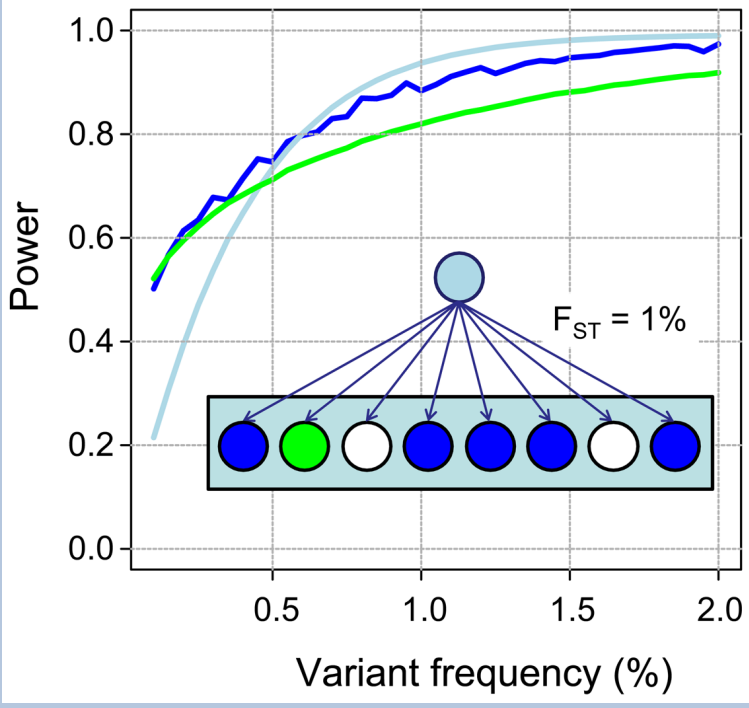

Figure.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Footnotes

[1] **Corresponding Author:** Author for correspondence Richard Durbin Wellcome Trust Sanger Centre Wellcome Trust Genome Campus, Hinxton, Cambridge. CB10 1HH UK Tel: +44 (0)1223 834244 Fax: +44 (0)1223 496802 rd@sanger.ac.uk

Steering Committee: David L. Altshuler(Co-Chair)23-4, Richard M. Durbin(Co-Chair)1, Gonçalo R. Abecasis5, David R. Bentley6, Aravinda Chakravarti7, Andrew G. Clark8, Francis S. Collins9, Francisco M. De La Vega10, Peter Donnelly11, Michael Egholm12, Paul Flicek13, Stacey B. Gabriel2, Richard A. Gibbs14, Bartha M. Knoppers15, Eric S. Lander2, Hans Lehrach16, Elaine R. Mardis17, Gil A. McVean11،18, Debbie A. Nickerson19, Leena Peltonen*, Alan J. Schafer20, Stephen T. Sherry21, Jun Wang22،23, Richard K. Wilson17
Sequencing Centres:Baylor College of Medicine Richard A. Gibbs (Principle Investigator)14, David Deiros14, Mike Metzker14, Donna Muzny14, Jeff Reid14, David Wheeler14 **BGI-Shenzhen** Jun Wang (Principle Investigator)22،23, Jingxiang Li22, Min Jian22, Guoqing Li22, Ruiqiang Li22،23, Huiqing Liang22, Geng Tian22, Bo Wang22, Jian Wang22, Wei Wang22, Huanming Yang22, Xiuqing Zhang22, Huisong Zheng22 **Broad Institute of MIT and Harvard** Eric S. Lander (Principal Investigator)2, David L. Altshuler23-4, Lauren Ambrogio2, Toby Bloom2, Kristian Cibulskis2, Tim J. Fennell2, Stacey B. Gabriel (Co-chair)2, Erica Shefler2, Carrie L. Sougnez2 **Illumina** David R. Bentley (Principle Investigator)6, Niall Gormley6, Sean Humphray6, Zoya Kingsbury6, Paula Koko-Gonzales6, Jennifer Stone6 **Life Technologies** Kevin J. McKernan (Principle Investigator)24, Gina L. Costa24, Jeffry K. Ichikawa24, Clarence C. Lee24 **Max Planck Institute for Molecular Genetics** Ralf Sudbrak (Project Leader)16, Hans Lehrach (Principal Investigator)16, Tatiana A. Borodina16, Andreas Dahl25, Alexey N. Davydov16, Peter Marquardt16, Florian Mertes16, Wilfiried Nietfeld16, Philip Rosenstiel26, Stefan Schreiber26, Aleksey V. Soldatov16, Bernd Timmermann16, Marius Tolzmann16 **Roche Applied Science** Michael Egholm (Principle Investigator)12, Jason Affourtit27, Dana Ashworth27, Said Attiya27, Melissa Bachorski27, Eli Buglione27, Adam Burke27, Amanda Caprio27, Christopher Celone27, Shauna Clark27, David Conners27, Brian Desany27, Lisa Gu27, Lorri Guccione27, Kalvin Kao27, Andrew Kebbel27, Jennifer Knowlton27, Matthew Labrecque27, Louise McDade27, Craig Mealmaker27, Melissa Minderman27, Anne Nawrocki27, Faheem Niazi27, Kristen Pareja27, Ravi Ramenani27, David Riches27, Wanmin Song27, Cynthia Turcotte27, Shally Wang27 **Washinton University in St. Louis** Elaine R. Mardis (Co-Chair) (Co-Principle Investigator) 17, Richard K. Wilson (Co-Principle Investigator)17, David Dooling17, Lucinda Fulton17, Robert

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

Fulton17, George Weinstock17 **Wellcome Trust Sanger Institute** Richard M. Durbin (Principle Investigator)1, John Burton1, David M. Carter1, Carol Churcher1, Alison Coffey1, Anthony Cox1, Aarno Palotie1, Michael Quail1, Tom Skelly1, James Stalker1, Harold P. Swerdlow1, Daniel Turner1
Analysis Group: Agilent Technologies Anniek De Witte28, Shane Giles28 **Baylor College of Medicine** Richard A. Gibbs (Principle Investigator)14, David Wheeler14, Matthew Bainbridge14, Danny Challis14, Aniko Sabo14, Fuli Yu14, Jin Yu14 **BGI-Shenzhen** Jun Wang (Principle Investigator)22،23, Xiaodong Fang22, Xiaosen Guo22, Ruiqiang Li22،23, Yingrui Li22, Ruibang Luo22, Shuaishuai Tai22, Honglong Wu22, Hancheng Zheng22, Xiaole Zheng22, Yan Zhou22, Guoqing Li22, Jian Wang22, Huanming Yang22 **Boston College** Gabor T. Marth (Principle Investigator)29, Erik P. Garrison29, Weichun Huang30, Amit Indap29, Deniz Kura29l, Wan-Ping Lee29, Wen Fung Leong29, Aaron R. Quinlan31, Chip Stewart29, Michael P. Stromberg6, Alistair N. Ward29, Jiantao Wu29 **Brigham and Women's Hospital** Charles Lee (Principle Investigator)32, Ryan E. Mills32, Xinghua Shi32 **Broad Institute of MIT and Harvard** Mark J. Daly (Principle Investigator)2, David L. Altshuler23¯4, Aaron D. Ball2, Eric Banks2, Toby Bloom2, Brian L. Browning33, Kristian Cibulskis2, Mark A. DePristo2, Tim J. Fennell2, Kiran V. Garimella2, Sharon R. Grossman2،34, Robert E. Handsaker2, Matt Hanna2, Chris Hartl2, Andrew M. Kernytsky2, Joshua M. Korn2, Heng Li2, Jared R. Maguire2, Steve A. McCarroll2, Aaron McKenna2, James C. Nemesh2, Anthony A. Philippakis2, Ryan E. Poplin2, Manuel A. Rivas2, Pardis C. Sabeti2،34, Stephen F. Schaffner2, Erica Shefler2, Ilya A. Shlyakhter2،34 Cardiff University, The Human Gene Mutation Database David N. Cooper (Principle Investigator)35, Edward V. Ball35, Matthew Mort35, Andrew D. Phillips35, Peter D. Stenson35 **Cold Spring Harbor Laboratory** Jonathan Sebat (Principle Investigator)36, Vladimir Makarov37, Kenny Ye38, Seungtai C. Yoon39 **Cornell and Stanford Universities** Carlos D. Bustamante (Co-Principle Investigator)40, Andrew G. Clark (Co-Principle Investigator)8, Alon Keinan (Co-Principle Investigator)8, Michael Snyder (Co-Principle Investigator)40, Adam Boyko40, Jeremiah Degenhardt8, Simon Gravel40, Fabian Grubert40, Ryan N. Gutenkunst41, Mark Kaganovich40, Phil Lacroute40, Xin Ma8, Andy Reynolds8, Alexander Urban40 **European Bioinformatics Institute** Laura Clarke (Project Leader)13, Paul Flicek (Co-Chair, DCC) (Principle Investigator)13, Fiona Cunningham13, Javier Herrero13, Stephen Keenen13, Eugene Kulesha13, Rasko Leinonen13, William M. McLaren13, Rajesh Radhakrishnan13, Richard E. Smith13, Vadim Zalunin13, Xiangqun Zheng-Bradley13 **European Molecular Biology Laboratory** Jan O. Korbel (Principle Investigator)42, Adrian M. Stütz42 **Illumina** Sean Humphray (Project Leader)6, Markus Bauer6, R. Keira Cheetham6, Tony Cox6, Michael Eberle6, Terena James6, Scott Kahn6, Lisa Murray6 **Johns Hopkins University** Aravinda Chakravarti7 **Leiden University Medical Center** Kai Ye43 **Life Technologies** Francisco M. De La Vega (Principle Invesigator)10, Yutao Fu24, Fiona C.L. Hyland10, Jonathan M. Manning24, Stephen F. McLaughlin24, Heather E. Peckham24, Onur Sakarya10, Yongming A. Sun10, Eric F. Tsung24 **Louisiana State University** Mark A. Batzer (Principle Investigator)44, Miriam K. Konkel44, Jerilyn A. Walker44 **Max Planck Institute for Molecular Genetics** Ralf Sudbrak (Project Leader)16, Marcus W. Albrecht16, Vyacheslav S. Amstislavskiy16, Ralf Herwig16, Dimitri Parkhomchuk16 **US National Institutes of Health** Stephen T. Sherry (Co-Chair, DCC) (Principle Investigator)21, Richa Agarwala21, Hoda M. Khouri21, Aleksandr O. Morgulis21, Justin E. Paschall21, Lon D. Phan21, Kirill E. Rotmistrovsky21, Robert D. Sanders21, Martin F. Shumway21, Chunlin Xiao21 **Oxford University** Gil A. McVean (Co-Chair) (Co-Chair, Population Genetics) (Principle Investigator)11،18, Adam Auton11, Zamin Iqbal11, Gerton Lunter11, Jonathan L. Marchini11،18, Loukas Moutsianas18, Simon Myers11،18, Afidalina Tumian18 **Roche Applied Science** Brian Desany (Project Leader)27, James Knight27, Roger Winer27 **The Translational Genomics Research Institute** David W. Craig (Principle Investigator)45, Steve M. Beckstrom-Sternberg45, Alexis Christoforides45, Nils Homer46, Ahmet A. Kurdoglu45, James O. Long45, Barry Merriman46, Stan F. Nelson46, John V. Pearson45, Shripad A. Sinari45, Waibhav D. Tembe45 **University of California, Santa-Cruz** David Haussler (Principle Investigator)47, Angie S. Hinrichs47, Sol J. Katzman47, Andrew Kern47, Robert M. Kuhn47 **University of Chicago** Molly Przeworski (Co-Chair, Population Genetics) (Principle Investigator)48, Ryan D. Hernandez49, Bryan Howie50, Joanna L. Kelley50, John C. Marioni50, S. Cord Melton50 **University of Michigan** Gonçalo R. Abecasis (Co-Chair) (Principle Investigator)5, Hyun M. Kang (Project Leader) 5, Paul Anderson5, Tom Blackwell5, Wei Chen5, William O. Cookson51, Jun Ding5, Mark Lathrop52, Yun Li5, Liming Liang53, Miriam F. Moffatt51, Paul Scheet54, Carlo Sidore5, Matthew Snyder5, Xiaowei Zhan5, Sebastian Zoellner5 **University of Montreal** Philip Awadalla (Principle Investigator)55, Ferran Casals56, Youssef Idaghdour56, John Keebler56, Eric A. Stone56, Martine Zilversmit56 **University of Utah** Lynn Jorde (Principle Investigator)57, Jinchuan Xing57 **University of Washington** Evan E. Eichler (Principle Investigator)58, Can Alkan58, Iman Hajirasouliha59, Fereydoun Hormozdiari59, Jeffrey M. Kidd40, S. Cenk Sahinalp59, Peter H. Sudmant19 **Washington University in St. Louis** Elaine R. Mardis (Principle Investigator)17, Ken Chen17, Asif Chinwalla17, Li Ding17, Daniel C. Koboldt17, Mike D. McLellan17, David Dooling17, George Weinstock17, John W. Wallis17, Michael C. Wendl17, Qunyuan Zhang17 **Wellcome Trust Sanger Institute** Richard M. Durbin (Principle Investigator)1, Cornelis Albers60, Qasim Ayub1, Senduran Balasubramaniam1, Jeffrey C. Barrett1, David M. Carter1, Yuan Chen1, Donald F. Conrad1, Petr Danecek1, Emmanouil T. Dermitzakis61, Min Hu1, Ni Huang1, Matt E. Hurles1, Hanjun Jin62, Luke Jostins1, Thomas M. Keane1, Quang Le1, Sarah Lindsay1, Quan Long1, Daniel G. MacArthur1, Stephen B. Montgomery61, Leopold Parts1, James Stalker1, Chris Tyler-Smith1, Klaudia Walter1, Yali Xue1, Bryndis, Yngvadottir1, Yujun Zhang1 **Yale University** Mark B. Gerstein (Principle Investigator)63،64, Alexej Abyzov63, Suganthi Balasubramanian65, Robert Bjornson64, Jiang Du64, Lukas Habegger63, Rajini Haraksingh63, Justin Jee63, Ekta Khurana65, Hugo Y.K. Lam40, Jing Leng63, Xinmeng Jasmine Mu63, Zhengdong Zhang65
Structural Variation Group: BGI-Shenzhen Yingrui Li22 **Boston College** Gabor T. Marth (Principle Investigator)29, Erik P. Garrison29, Deniz Kura29l, Aaron R. Quinlan31, Chip Stewart29, Michael P. Stromberg6, Alistair N. Ward29, Jiantao Wu29 **Brigham and Women's Hospital** Charles Lee (Co-Chair) (Principle Investigator)32, Ryan E. Mills32, Xinghua Shi32 **Broad Institute of MIT and Harvard** Eric Banks2, Mark A. DePristo2, Robert E. Handsaker2, Chris Hartl2, Joshua M. Korn2, Heng Li2, Steve A. McCarroll2, James C.

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

Nemesh2 **Cold Spring Harbor Laboratory** Jonathan Sebat (Principle Investigator)36, Vladimir Makarov37, Kenny Ye38, Seungtai C. Yoon39 **Cornell and Stanford Universities** Michael Snyder (Co-Principle Investigator)40, Jeremiah Degenhardt8, Fabian Grubert40, Mark Kaganovich40, Alexander Urban40 **European Bioinformatics Institute** Laura Clarke (Project Leader)13, Richard E. Smith13, Xiangqun Zheng-Bradley13 **European Molecular Biology Laboratory** Jan O. Korbel42 **Illumina** Sean Humphray (Project Leader)6, R. Keira Cheetham6, Michael Eberle6, Scott Kahn6, Lisa Murray6 **Leiden University Medical Center** Kai Ye43 **Life Technologies** Francisco M. De La Vega (Principle Invesigator)10, Yutao Fu24, Heather E. Peckham24, Yongming A. Sun10 **Louisiana State University** Mark A. Batzer (Principle Investigator)44, Miriam K. Konkel44, Jerilyn A. Walker44 **US National Institutes of Health** Chunlin Xiao21 **Oxford University** Zamin Iqbal11 **Roche Applied Science** Brian Desany27 **University of Michigan** Tom Blackwell (Project Leader)5, Matthew Snyder5 **University of Utah** Jinchuan Xing57 **University of Washington** Evan E. Eichler (Principle Investigator)58, Can Alkan58, Iman Hajirasouliha59, Fereydoun Hormozdiari59, Jeffrey M. Kidd40 **Washington University in St. Louis** Ken Chen17, Asif Chinwalla17, Li Ding17, Mike D. McLellan17, John W. Wallis17 **Wellcome Trust Sanger Institute** Matt E. Hurles1 (Co-Chair), Donald F. Conrad1, Klaudia Walter1, Yali Xue1, Yujun Zhang1 **Yale University** Mark B. Gerstein (Principle Investigator)63‚64, Alexej Abyzov63, Jiang Du64, Rajini Haraksingh63, Justin Jee63, Ekta Khurana65, Hugo Y.K. Lam40, Jing Leng63, Xinmeng Jasmine Mu63, Zhengdong Zhang65
Exon Pilot Group: Baylor College of Medicine Richard A. Gibbs (Co-Chair) (Principle Investigator)14, Matthew Bainbridge14, Danny Challis14, Cristian Coafra14, Huyen Dinh14, Christie Kovar14, Sandy Lee14, Donna Muzny14, Lynne Nazareth14, Jeff Reid14, Aniko Sabo14, Fuli Yu14, Jin Yu14 **Boston College** Gabor T. Marth (Co-Chair) (Principle Investigator)29, Erik P. Garrison29, Amit Indap29, Wen Fung Leong29, Aaron R. Quinlan31, Chip Stewart29, Alistair N. Ward29, Jiantao Wu29 **Broad Institute of MIT and Harvard** Kristian Cibulskis2, Tim J. Fennell2, Stacey B. Gabriel2, Kiran V. Garimella2, Chris Hartl2, Erica Shefler2, Carrie L. Sougnez2, Jane Wilkinson2 **Cornell and Stanford Universities** Andrew G. Clark (Co-Principle Investigator)8, Simon Gravel40, Fabian Grubert40 **European Bioinformatics Institute** Laura Clarke (Project Leader)13, Paul Flicek (Co-Chair, DCC) (Principle Investigator)13, Richard E. Smith13, Xiangqun Zheng-Bradley13 **US National Institutes of Health** Stephen T. Sherry (Co-Chair, DCC) (Principle Investigator)21, Hoda M. Khouri21, Justin E. Paschall21, Martin F. Shumway21, Chunlin Xiao21 **Oxford University** Gil A. McVean (Principle Investigator)11‚18, Zamin Iqbal11 **University of California, Santa-Cruz** Sol J. Katzman47 **University of Michigan** Gonçalo R. Abecasis (Co-Chair) (Principle Investigator)5, Tom Blackwell5 **University of Washington** Debbie A. Nickerson19, Peter H. Sudamant19 **Washington University in St. Louis** Elaine R. Mardis (Principle Investigator)17, David Dooling17, Lucinda Fulton17, Robert Fulton17, Daniel C. Koboldt17 **Wellcome Trust Sanger Institute** Richard M. Durbin (Principle Investigator)1, Senduran Balasubramaniam1, Allison Coffey1, Thomas M. Keane1, Daniel G. MacArthur1, Aarno Palotie1, Carol Scott1, James Stalker1, Chris Tyler-Smith1 **Yale University** Mark B. Gerstein (Principle Investigator)63‚64, Suganthi Balasubramanian65
Samples and ELSI Group: Aravinda Chakravarti (Co-chair)7, Bartha Knoppers (Co-chair)15, Leena Peltonen (Co-chair)*, Gonçalo Abecasis5, Carlos D. Bustamante40, Neda Gharani66, Richard Gibbs14, Lynn Jorde57, Jane Kaye67, Alastair Kent68, Taosha Li22, Amy McGuire69, Gil McVean11‚18, Pilar Ossorio70, Charles Rotimi71, Yeyang Su22, Lorraine Toji66, Chris Tyler-Smith1, Huanming Yang22
Scientific Management: Lisa Brooks72, Adam L. Felsenfeld72, Jean McEwen72, Assya Abdallah73, Christopher Juenger74, Francis Collins9, Audrey Duncanson20, Eric Green75, Mark Guyer72, Jane Peterson72, Alan Schafer20
Writing Group: Gonçalo Abecasis5, Adam Auton11, David Altshuler23¯4, Lisa Brooks72, Richard Durbin1, Richard Gibbs14, Matt Hurles1, Gil McVean11‚18

Competing Financial Interest Statements A.C. is on the Scientific Advisory Board of Affymetrix, Inc.; E.E.E. is a member of the Scientific Advisory Board for Pacific Biosciences; A.L.M. advises Ion Torrents Systems; M.S. is a member of the Scientific Advisory Boards of DNANexus and GenapSis; M.B, D.R.B., R.K.C., T.C., M.E., N.G., S.H., T.J., S.K., Z.K. P.K-G., L.M., J.S., & M.P.S. work for Illumina Cambridge Ltd.; G.L.C., F.M.D.L.V., Y.F., F.C.L.H., J.K.I., C.C.L., J.M.M., K.J.M., S.F.M., H.E.P., O.S., Y.A.S., & E.F.T. work for Life Technologies; J.A., D.A., S.A., M.B., E.B., A.B., A.C., C.C., S.C., D.C., B.D., M.E., L.G., L.G., K.K., A.K., J.K., J.K., M.L., L.M., C.M., M.M., A.N., F.N., K.P., R.R., D.R., W.S., C.T., S.W., & R.W. work for Roche Applied Science.

Acknowledgments

We thank many people who contributed to this project: K. Beal, S. Fitzgerald, G. Cochrane, V. Silventoinen, P. Jokinen, and E. Birney (European Bioinformatics Institute); T Hunkapiller and Q. Doan (Life Technologies) for their advice and coordination; N. Kälin (German Aerospace Center [DLR]); F. Laplace (German Federal Ministry of Education and Research [BMBF]); J. Wilde, S. Paturej, and I. Kühndahl (Max Planck Institute for Molecular Genetics); J. Knight and C. Kodira (Roche Applied Science); M. Boehnke (University of Michigan) for valuable discussions; Z. Cheng, S. Sajjadian, and F. Hormozdiari (University of Washington) for assistance in managing datasets; J.Ahringer (University of Cambridge) for comments on the manuscript; D. Leja (NIH) for help with the figures; and N. Clemm (NIH).

We thank the Yoruba in Ibadan, Nigeria, the Han Chinese in Beijing, China, the Japanese in Tokyo, Japan, the Utah CEPH community, the Luhya in Webuye, Kenya, the Toscani in Italia, and the Chinese in Denver, Colorado, for contributing samples for research.

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

This research was supported in part by Wellcome Trust grants WT085532AIA to P.F. and WT086084/Z/08/Z to G.A.M.; WT081407/Z/06/Z to J.S.K.; WT075491/Z/04 to G.L.; WT077009 to C.T-S.; Medical Research Council G0801823 to J.L.M.; British Heart Foundation grant RG/09/012/28096 to C.A.; The Leverhulme Trust and EPSRC studentships to L.M. and A.T.; the Louis-Jeantet Foundation and Swiss National Science Foundation in support of S.B.M.; NGI/EBI fellowship 050-72-436 to K.Y.; National Natural Science Foundation of China grants 30725008, 30890032, 30811130531, and 30221004; the Chinese 863 program grants 2006AA02Z177, 2006AA02Z334, 2006AA02A302, and 2009AA022707; 973 Program grant S2010081019; the Municipal Government of Shenzhen, China grants JC200903190767A, JC200903190772A, ZYC200903240076A, CXB200903110066A, ZYC200903240077A, ZYC200903240076A, and ZYC200903240080A; the Yantian District local government of Shenzhen; the Ole Rømer grant from the Danish Natural Science Research Council; an Emmy Noether Fellowship of the German Research Foundation (Deutsche Forschungsgemeinschaft) to J.O.K.; BMBF grant 01GS08201; BMBF grant PREDICT 0315428A to R.H.; BMBF NGFN PLUS and EU 6th framework READNA to S.S.; EU 7th framework 242257 to A.V.S.; the Max Planck Society; a grant from Genome Quebec and the Ministry of Development, Exploration and Innovation PSR-SIIRI-195 to P.A.; the Intramural Research Program of the NIH, the National Library of Medicine and NIH grants HG4221 and HG5209 to C.L.; HG4222 to J.S.; GM59290 to L.B.J. and M.A.B.; GM72861 to M.P.; HG2651 and MH84698 to G.R.A.; HG5214 to G.R.A. and A.C.; HG4120 to E.E.E.; HG2750 to D.L.A.; HG2757 to A.C.; HG2510 to D.C.; HG5208 to M.J.D.; HG3273 and HG5211 to R.A.G.; HG3698, HG4719, and HG5552 to G.M.; HG3229 to C.D.B. and A.G.C.; HG2357 to M.S.; HG3067 to E.B.; HG4960 to B.L.B; HG2371 and HG4568 to D.H.; HG5201 to A.A.K.; HG3698 and HG4719 to G.T.M.; HG4333 to A.M.L.; N01-HG-62088 to the Coriell Institute; HG5210 to the Translational Genomics Research Institute; Al Williams Professorship funds for M.B.G.; the BWF and Packard Foundation support for P.C.S.; and the Pew Charitable Trusts support for G.R.A.; a NSF Minority Postdoctoral Fellowship in support of R.D.H.; E.E.E. is an HHMI investigator and M.P. is an HHMI Early Career Scientist.

References

1. The International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 2004;431:931–45. [PubMed: 15496913]
2. Sachidanandam R, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 2001;409:928–33. [PubMed: 11237013]
3. The International HapMap Consortium. A haplotype map of the human genome. Nature 2005;437:1299–320. [PubMed: 16255080]
4. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007;449:851–61. [PubMed: 17943122]
5. NHGRI Office of Population Genomics. A catalog of published genome-wide association studies. 2010
6. Craddock N, et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 2010;464:713–20. [PubMed: 20360734]
7. Manolio TA, et al. Finding the missing heritability of complex diseases. Nature 2009;461:747–53. [PubMed: 19812666]
8. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 2009;324:387–9. [PubMed: 19264985]
9. Cohen JC, Boerwinkle E, Mosley TH Jr. Hobbs HH. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N Engl J Med 2006;354:1264–72. [PubMed: 16554528]
10. Levy S, et al. The diploid genome sequence of an individual human. PLoS Biol 2007;5:e254. [PubMed: 17803354]
11. Wheeler DA, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature 2008;452:872–6. [PubMed: 18421352]
12. Bentley DR, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008;456:53–9. [PubMed: 18987734]
13. Wang J, et al. The diploid genome sequence of an Asian individual. Nature 2008;456:60–5. [PubMed: 18987735]
14. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. Nature. 2010 In press.
15. Liti G, et al. Population genomics of domestic and wild yeasts. Nature 2009;458:337–41. [PubMed: 19212322]

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

16. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet 2009;10:387–406. [PubMed: 19715440]
17. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078–9. [PubMed: 19505943]
18. Albers C, et al. Dindel: Accurate indel calls from short read data. Genome Res. 2010 submitted.
19. Lam HY, et al. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. Nat Biotechnol 2010;28:47–55. [PubMed: 20037582]
20. Conrad DF, et al. Origins and functional impact of copy number variation in the human genome. Nature 2010;464:704–12. [PubMed: 19812545]
21. Irwin JA, et al. Investigation of heteroplasmy in the human mitochondrial DNA control region: a synthesis of observations from more than 5000 global population samples. J Mol Evol 2009;68:516–27. [PubMed: 19407924]
22. Balaresque P, et al. A predominantly neolithic origin for European paternal lineages. PLoS Biol 2010;8:e1000285. [PubMed: 20087410]
23. Wendl MC, Wilson RK. The theory of discovering rare variants via DNA sequencing. BMC Genomics 2009;10:485. [PubMed: 19843339]
24. Quang LS, Li H, Durbin R. QCALL: A genealogical method for variant detection and genotyping from low coverage sequence data in population samples. Genome Res. submitted.
25. Seattle SNPs. NHLBI Program for Genomic Applications. 2010 SeattleSNPs.
26. Xing J, et al. Mobile elements create structural variation: analysis of a complete human genome. Genome Res 2009;19:1516–26. [PubMed: 19439515]
27. Stranger BE, et al. Population genomics of human gene expression. Nat Genet 2007;39:1217–24. [PubMed: 17873874]
28. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epi 2010;32:1–19.
29. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet 2010;11:499–511. [PubMed: 20517342]
30. Dixon AL, et al. A genome-wide association study of global gene expression. Nature Genetics 2007;39:1202–07. [PubMed: 17873877]
31. Genovese G. Association of Trypanolytic ApoL1 Variants with Kidney Disease in African-Americans. Science. 2010
32. Nachman MW, Crowell SL. Estimate of the mutation rate per nucleotide in humans. Genetics 2000;156:297–304. [PubMed: 10978293]
33. Kondrashov AS. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. Hum Mutat 2003;21:12–27. [PubMed: 12497628]
34. Roach JC, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 2010;328:636–9. [PubMed: 20220176]
35. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics 1993;134:1289–303. [PubMed: 8375663]
36. Maynard Smith J, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res 1974;23:23–35. [PubMed: 4407212]
37. Cai JJ, Macpherson JM, Sella G, Petrov DA. Pervasive hitchhiking at coding and regulatory sites in humans. PLoS Genet 2009;5:e1000336. [PubMed: 19148272]
38. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol 2006;4:e72. [PubMed: 16494531]
39. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. Natural selection has driven population differentiation in modern humans. Nat Genet 2008;40:340–5. [PubMed: 18246066]
40. Lamason RL, et al. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science 2005;310:1782–6. [PubMed: 16357253]
41. Tournamille C, Colin Y, Cartron JP, Le Van Kim C. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. Nat Genet 1995;10:224–8. [PubMed: 7663520]

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

42. Myers S, et al. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science 2010;327:876–9. [PubMed: 20044541]

43. Myers S, Freeman C, Auton A, Donnelly P, McVean G. A common sequence motif associated with recombination hot spots and genome instability in humans. Nat Genet 2008;40:1124–9. [PubMed: 19165926]

44. Baudat F, et al. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science 2010;327:836–40. [PubMed: 20044539]

45. Parvanov ED, Petkov PM, Paigen K. Prdm9 controls activation of mammalian recombination hotspots. Science 2010;327:835. [PubMed: 20044538]

46. Liu JZ, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat Genet 2010;42:436–40. [PubMed: 20418889]

47. Sanna S, et al. Variants within the immunoregulatory CBLB gene are associated with multiple sclerosis. Nat Genet 2010;42:495–7. [PubMed: 20453840]

48. Musunuru K, Kathiresan S. Exome Sequencing Identifies ANGPTL3 as a Cause of Familial Combined Hypolipidemia. New Engl. J. Med. In review.

49. Ewing AD, Kazazian HH Jr. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res. 2010

50. Mills RE, et al. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res 2006;16:1182–90. [PubMed: 16902084]

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

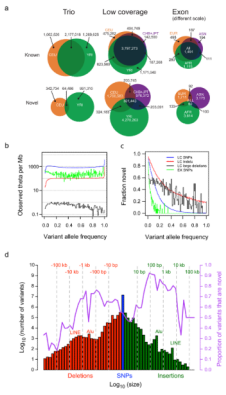

Figure 1. Properties of the variants found

a, Venn diagrams showing the numbers of SNPs identified in each pilot project in each population or analysis panel, subdivided according to whether the SNP was present in dbSNP release 129 (“Known”) or not (“Novel”). Exon analysis panel AFR is YRI+LWK, ASN is CHB+CHD+JPT, and EUR is CEU+TSI. Note that the scale for the exon project column is much larger than for the other pilots. **b,** The number of variants per Mb at different allele frequencies divided by the expectation under the neutral coalescent (1/*i*, where *i* is the variant allele count), thus giving an estimate of theta per megabase. Blue: low coverage SNPs, red: low coverage indels, black: low coverage genotyped large deletions, green: exon SNPs. The spikes at the right ends of the lines correspond to excess variants for which all samples differed from the reference (approximately 1 per 30 kb), consistent with errors in the reference sequence. **c**, Fraction of variants in each allele frequency class that were novel. Novelty was determined by comparison to dbSNP release 129 for SNPs and small indels, dbVar (June 2010) for deletions, and two published genomes[10, 11] for larger indels. **d,** Size distribution and novelty of variants discovered in the low coverage project. SNPs are shown in blue, deletions with respect to the reference sequence in red, and insertions or duplications with respect to the reference in green. The fraction of variants in each size bin that were novel is shown by the purple line, and is defined relative to dbSNP (SNPs and indels), dbVar (deletions, duplications, mobile element insertions), dbRIP and other studies[49] (mobile element insertions), Venter and Watson genomes[10, 11] (indels and deletions), and indels from split capillary reads[50] (indels and deletions). To account for ambiguous placement of many indels, discovered indels were deemed to match known indels if they were within 25 bp of a known indel of the same size. To account for imprecise knowledge of the location of most deletions and duplications, discovered variants were deemed to match known variants if they had > 50% reciprocal overlap.

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

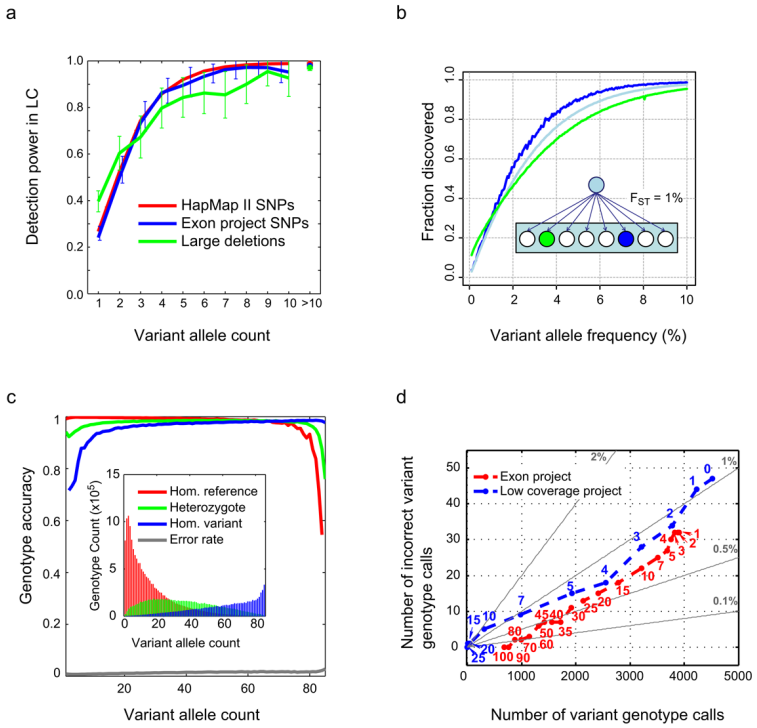

Figure 2. Variant discovery rates and genotype accuracy in the low coverage project

a, Rates of low coverage variant detection by allele frequency in CEU. Lines show the fraction of variants seen in overlapping samples in independent studies, that were also found to be polymorphic in the low coverage project (in the same overlapping samples), as a function of allele count in the 60 low coverage samples. Note that we plot power against expected allele count in 60 samples, e.g. a variant present in, say, 2 copies in an overlap of 30 samples is expected to be present 4 times in 60 samples. The crosses on the right represent the average discovery fraction for all variants having more than 10 copies in the sample. Colours correspond to: (red) HapMap II sites, excluding sites also in HapMap 3 (43 overlapping samples); (blue) exon project sites (57 overlapping samples); (green) deletions from Conrad et al.[20] (60 overlapping samples; deletions were classified as "found" if there was any overlap). **b,** Estimated rates of discovery of variants at different frequencies in the CEU (blue), a population related to the CEU with Fst = 1% (green) and across Europe as a whole (light blue). The insert shows a cartoon of the statistical model for population history and thus allele frequencies in related populations where an ancestral population gave rise to many equally related populations, one of which (blue circle) has samples sequenced. **c,** SNP genotype accuracy by allele frequency in the CEU low coverage project, measured by comparison to HapMap II genotypes at sites present in both call sets, excluding sites that were also in HapMap 3. Lines represent the average accuracy of homozygote reference (red), heterozygote (green) and homozygote alternative calls (blue) as a function of the alternative allele count in the overlapping set of 43 samples, and the overall genotype error rate (grey, at bottom of plot). The inset shows the number of each genotype class as a function of alternative allele count. **d,** Coverage and accuracy for the low coverage and exon projects as a function of depth threshold. For 41 CEU samples sequenced in both the exon and low coverage projects, on the x axis is shown the number of non-reference SNP genotype calls at HapMap II sites not in HapMap 3 that were called in the exon project target region, and on the y axis is shown the number of these calls that were not variant (i.e., are reference homozygote and thus incorrectly were called as variant) according to HapMap

II. Each point plotted corresponds to a minimum depth threshold for called sites. Grey lines show constant error rates. The exon project calls (red) were made independently per sample, whereas the low coverage calls (blue), which were only slightly less accurate, were made using LD information that combined partial information across samples and sites in an imputation-based algorithm. The additional data added from point “1” to point “0” (upper right in the figure) for the low coverage project were completely imputed.

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

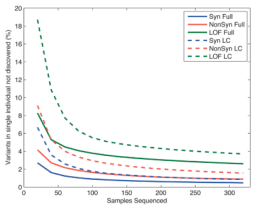

Figure 3. The value of additional samples for variant discovery

The fraction of variants present in an individual that would not have been found in a sequenced reference panel, as a function of reference panel size and the sequencing strategy. The lines represent predictions for Synonymous (Syn), Nonsynonymous (NonSyn), and Loss of function (LOF) variant classes, broken down by sequencing category: full sequencing as for exons (Full) and low coverage sequencing (LowCov). The values were calculated from observed distributions of variants of each class in 321 East Asian samples (CHB, CHD and JPT populations) in the exon data, and power to detect variants at low allele counts in the reference panel from Figure 2a.

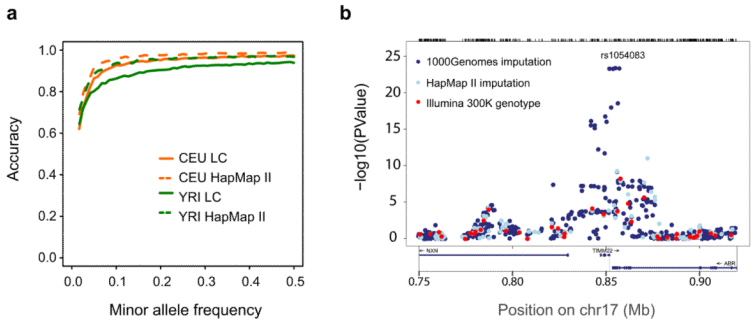

Figure 4. Imputation from the low coverage data

a, Accuracy of imputing variant genotypes using HapMap 3 sites to impute sites from the low coverage (LC) project into the trio fathers as a function of allele frequency. Accuracy of imputing genotypes from the HapMap II reference panels[4] is also shown. Imputation accuracy for common variants was generally a few percent worse from the low coverage project than from HapMap, although error rates increase for less common variants. **b,** An example of imputation in a cis-eQTL for *TIMM22*, for which the original Ilumina 300K genotype data gave a weak signal[30]. Imputation using HapMap data made a small improvement, and imputation using low coverage haplotypes provided a much stronger signal.

UKPMC Funders Group Author Manuscript
UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

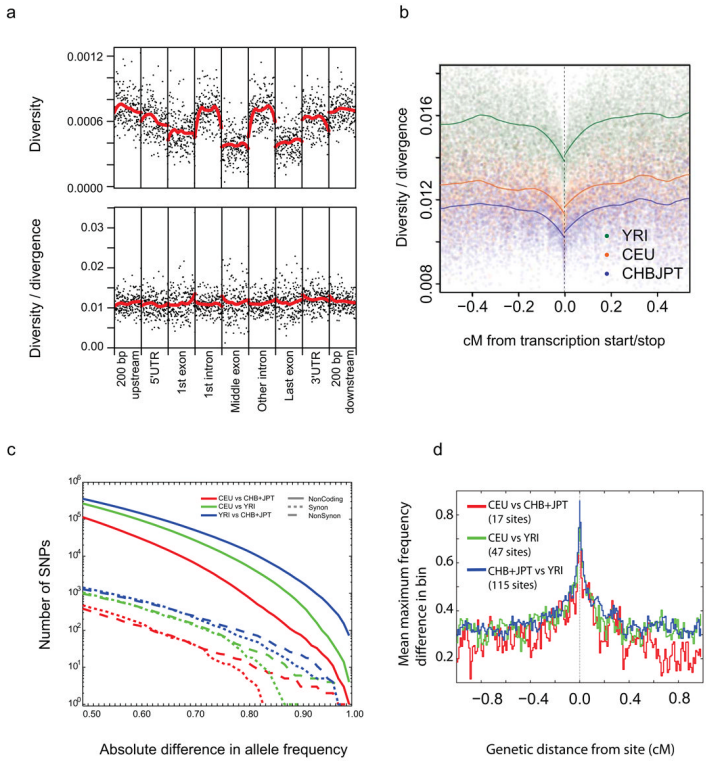

Figure 5. Variation around genes

a, Diversity in genes calculated from the CEU low coverage genotype calls (upper) and diversity divided by divergence between humans and rhesus macaque (lower). Within each element averaged diversity is shown for the first and last 25 base pairs, with the remaining 150 positions sampled at fixed distances across the element (elements shorter than 150 base pairs were not considered). Note that estimates of diversity will be reduced compared to the true population value due to the reduced power for rare variants, but relative values should be little affected. **b,** Average autosomal diversity divided by divergence, as a function of genetic distance from coding transcripts, calculated at putatively neutral sites, i.e., excluding phastcons conserved noncoding sequences and all sites in coding exons but four-fold degenerate sites. **c,** Numbers of SNPs showing increasingly high levels of differentiation in allele frequency between the CEU and CHB+JPT (red), CEU and YRI (green) and CHB +JPT and YRI (blue). Lines indicate synonymous variants (dashed), nonsynonymous variants (dotted) and other variants (solid). The most highly differentiated genic SNPs were enriched for nonsynonymous variants, indicating local adaptation. **d,** The decay of population differentiation around genic SNPs showing extreme allele frequency differences between populations (difference in frequency of at least 0.8 between populations, thinned so there is no more than one per gene considered). For all such SNPs the highest allele frequency difference in bins of 0.01 cM away from the variant was recorded and averaged.

UKPMC Funders Group Author Manuscript

UKPMC Funders Group Author Manuscript

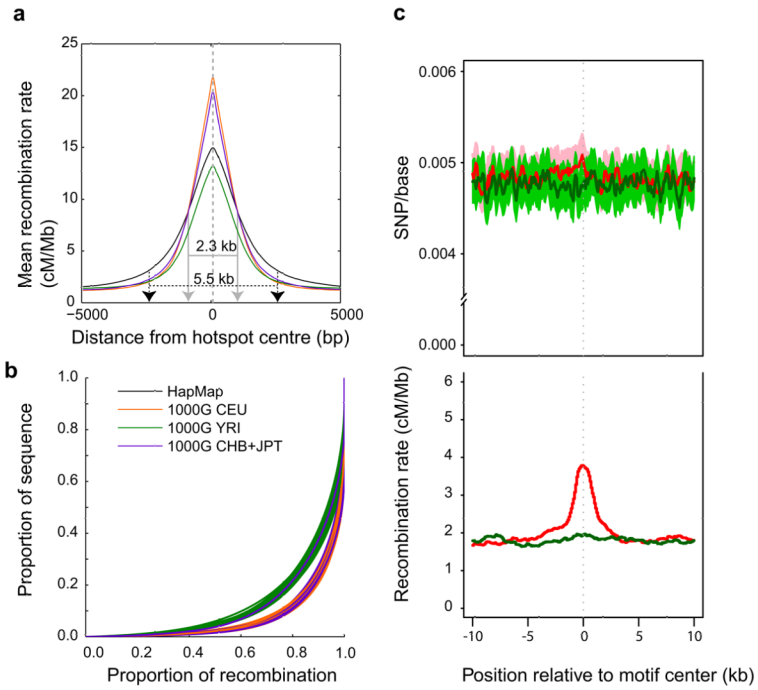

Figure 6. Recombination

a, Improved resolution of hotspot boundaries. The average recombination rate estimated from low coverage project data around recombination hotspots detected in HapMap II. Recombination hotspots were narrower, and in CEU (orange) and CHB+JPT (purple) more intense than previously estimated. **b,** The concentration of recombination in a small fraction of the genome, one line per chromosome. If recombination were uniformly distributed throughout the genome, then the lines on this figure would appear along the diagonal. Instead, most recombination occurs in a small fraction of the genome. Recombination rates in YRI (green) appeared to be less concentrated in recombination hotspots than CEU (orange) or CHB+JPT (purple). HapMap II estimates are shown in black. **c,** The relationship between genetic variation and recombination rates in the YRI population. The top plot shows average levels of diversity, measured as mean number of segregating sites per base, surrounding occurrences of the previously described hotspot motif [43] (CCTCCCTNNCCAC, red line) and a closely related, but not recombinogenic DNA sequence (CTTCCCTNNCCAC, green line). The lighter red and green shaded areas give 95% confidence intervals on diversity levels. The bottom plot shows estimated mean recombination rates surrounding motif occurrences, with colours defined as in the top plot.

UKPMC Funders Group Author Manuscript
UKPMC Funders Group Author Manuscript

Nature. Author manuscript; available in PMC 2011 April 1.

Table 1

Variants discovered by pilot, type, population and novelty.

	Low coverage				Trios			Exon	Union
	CEU	**YRI**	**CHB+JPT**	**Total**	**CEU**	**YRI**	**Total**	**Total**	**Total**
Samples	60	59	60	179	3	3	6	697	742
Total raw bases (Gb)	1401.56	874.40	595.93	2871.89	560.38	614.63	1175.01	845.40	4892.30
Total mapped bases (Gb)	817.46	595.58	468.17	1881.20	368.89	342.47	711.36	55.74	2648.29
Mean mapped depth (x)	4.62	3.42	2.65	3.56	43.14	40.05	41.60	55.92	N/A
Fraction of genome called	2.43 Gb (86%)	2.39 Gb (85%)	2.41 Gb (85%)	2.42 Gb (86.0%)	2.26 Gb (79%)	2.21 Gb (78%)	2.24 Gb (79%)	1.4 Mb	N/A
No. of SNPs (% novel)	7,943,827 (33%)	10,938,130 (47%)	6,273,441 (28%)	14,894,361 (54%)	3,646,764 (11%)	4,502,439 (23%)	5,907,699 (24%)	12,758 (70%)	15,275,256 (55%)
Variant SNP sites / individual	2,918,623	3,335,795	2,810,573	3,019,909	2,741,276	3,261,036	3,001,156	763	N/A
No. of indels (% novel)	728,075 (39%)	941,567 (52%)	666,639 (39%)	1,330,158 (57%)	411,611 (25%)	502,462 (37%)	682,148 (38%)	21 (52%)	1,480,877 (57%)
Variant indel sites / individual	354,767	383,200	347,400	361,669	322,078	382,869	352,474	1	N/A
No. of deletions (% novel)	N/D	N/D	N/D	15,893 (60%)	6,593 (41%)	8,129 (50%)	11,248 (51%)	N/D	22,025 (61%)
No. of genotyped deletions (% novel)	N/D	N/D	N/D	10,742 (57%)	N/D	N/D	6,317 (48%)	N/D	13,826 (58%)
No. of duplications (% novel)	259 (90%)	320 (90%)	280 (91%)	407 (89%)	187 (93%)	192 (91%)	256 (92%)	N/D	501 (89%)
No. of mobile element insertions (% novel)	3,202 (79%)	3,105 (84%)	1,952 (76%)	4,775 (86%)	1,397 (68%)	1,846 (78%)	2,531 (78%)	N/D	5,371 (87%)
No. of novel sequence insertions (% novel)	N/D	N/D	N/D	N/D	111 (96%)	66 (86%)	174 (93%)	N/D	174 (93%)

	Exon populations						
	CEU	**TSI**	**LWK**	**YRI**	**CHB**	**CHD**	**JPT**
Samples	90	66	108	112	109	107	105
Total collected bases (Gb)	151.15	63.96	53.42	146.52	93.08	126.59	210.68
Mean mapped depth on target (x)	73	71	32	62	47	62	53
No. of SNPs (% novel)	3,489 (34%)	3,281 (34%)	5,459 (50%)	5,175 (46%)	3,415 (47%)	3,431 (50%)	2,900 (42%)
Variant SNP sites / individual	715	727	902	794	713	770	694
No. of indels (No. novel)	8 (2)	8 (2)	N/D	10 (4)	9 (4)	4 (1)	9 (3)
Variant indel sites / individual	2	2	N/D	1	1	0	1

UKPMC Funders Group Author Manuscript
UKPMC Funders Group Author Manuscript

Nature. Author manuscript; available in PMC 2011 April 1.

Table 2

Estimated numbers of potentially functional variants in genes.

	Combined	Combined	Low Coverage		High-Coverage Trio		Exon Capture		
class	**Total**	**Novel**	**Total**	**Individual***	**Total**	**Individual***	**Total**	**Individual***	**GENCODE extrap.**
synonymous SNPs	60157	23498	55217	10572-12126	21410	9193-12500	5708	461-532	11553-13333
nonsynonymous SNPs	68300	34161	61284	9966-10819	19824	8299-10866	7063	396-441	9924-11052
small in-frame indels	714	383	666	198-205	289	130-178	20	1-3	~50-100
stop losses	77	40	71	9-11	22	4-14	6	0-0	~0
stop-introducing SNPs	1057	755	951	88-101	192	67-100	82	2-3	~50-75
splice-site-disrupting SNPs	517	399	500	41-49	82	28-45	3	1-1	~50
small frameshift indels	954	551	890	227-242	433	192-280	2	0	~0
genes disrupted by large deletions	147	71	143	28-36	82	33-49	NA	NA	NA
total genes containing LOF variants	2304	NA	1795	272-297	483	240-345	77	2-3	~50-75
HGMD "damaging mutation" SNPs	671	NA	578	57-80	161	48-82	99	2-4	~50-100

*Interquartile range of number of SNPs per individual.