Abbreviations

1000 Genomes Project (1000GP)
Structural variant (SV)
Mobile Element Locator Tool (MELT)
discordant read pairs (DPs)
split reads (SRs)
read-depth (RD)
paired-end (PE)
target site duplications (TSDs)
aCGH array comparative genomic hybridization
intensity rank sum (IRS)
median absolute deviation (MAD)
microhomology (MH)
nonhomology-based SV formation mechanism, such as nonhomologous end-joining or microhomology-mediated break-induced replication (NH)
digital comparative genomic hybridization (dCGH)
false discovery rate (FDR)
single nucleotide polymorphism (SNP)
insertions/deletions (indels)
copy number variant (CNV)
genome-wide association studies (GWAS)
linkage disequilibrium (LD)
logarithm of odds (LOD)
whole-genome sequencing (WGS)
circular consensus sequences (CCS) reads
continuous long reads (CLR)
copy number variable regions (CNVRs)
variant allele frequency (VAF)
principal component analysis (PCA)
reciprocal overlap (RO)
residual variation intolerance score (RVIS)
untranslated region (UTR)
Hardy-Weinberg equilibrium (HDW)
Coding sequence (CDS)
Intensity rank sum test (IRS test)
long runs of homozygosity (LROH)

Types of Structural Variation:

DEL	Deletion
DUP	Duplication
mCNV	Multi-allelic Copy Number Variant
INV	Inversion
MEI	Mobile Element Insertion
NUMT	Nuclear Mitochondrial Insertion

List of Supplementary Notes

1. Supplementary Text

1.1. DNA Samples

Structural variant (SV) discovery and genotyping was performed using 7.4-fold coverage Illumina paired-end population-scale sequencing data available at 1000genomes.org. Unless indicated otherwise, all SV discovery and genotyping algorithms were executed on the set of $N_{initial}$=2,535 individual samples initially designated for inclusion in phase3 (1000genomes.org). The final set of N_{final}=2,504 phase 3 samples, including lymphoblastoid cell line (N=2,400) as well as blood (N=104) DNA based samples, was obtained after removal of data from N=31 samples thereby avoiding inclusion of individuals exhibiting cryptic relationship patterns (1000genomes.org). Somatic SVs common to lymphoblast cell lines (*i.e.* regions undergoing site-specific somatic structural variation in B cells) were removed from the final callset (by removing all SVs from the chromosome regions shown below in Table 1.1.1). Within individual populations, we noticed only few differences in SV counts between sequenced DNA samples prepared from cell-line versus blood-derived DNAs. Cell lines/DNA for all samples (IDs provided in ED Table 1) are available from Coriell.

Chromosome	Start	End	ID
2	89156874	90471176	IGVDJ
7	38292981	38407770	TRVDJ
7	141999017	142511084	TRVDJ
9	33618463	33662656	TRVDJ
14	22089991	23014042	TRVDJ
14	105994256	107283085	IGVDJ
22	22385572	23265082	IGVDJ

Table 1.1.1. Regions of somatic rearrangement in lymphoblast removed from our SV map.

2. SV discovery & genotyping

2.1. BreakDancer

SV Type(s): DEL

Contributed by: Wanding Zhou, Zechen Chong, Xian Fan, Klaudia Walter, Ken Chen

Deletion callset generation: BreakDancer (BD, v1.1.2)[1] was run on all whole genome sequenced samples following BWA alignment. Deletion calls were made by chromosome and separately for each population using reads with mapping quality greater or equal to 20. Insert size distributions were analyzed for each library separately using a 1Mb region on chr20 (chr20:10000000-11000000) to determine thresholds which replaced the upper cutoffs (=upper) in the BD config files. The upper cutoff represents the upper boundary of the expected insert size distribution. To obtain a conservative estimate of the upper-cutoff, three different types of thresholds were calculated, (1) the drop in the density function of each insert size distribution, (2) the median plus four times the standard deviation, (3) the median plus five times the median absolute deviation (MAD); the maximum of those three estimates was chosen. About 1% of the libraries showed an extreme insert size distribution, whereby either the median insert size or the third quantile of the insert size distribution was zero (only insert sizes ≥0 were extracted from the bam files), in those cases the cutoffs 1,000 and 10,000 respectively were chosen.

The raw BD calls were filtered for deletion size (< 50 bp and >1 Mbp), for estimated read-depth ratio (< 0.75), for number of spanning read pairs (≥20), for regions around centromeres (+/- 1 kbp), for regions around assembly gaps (+/- 50 bp) and for alpha satellite regions. The read-depth (RD) ratio was calculated as the average RD of the samples that supported the deletion divided by the average RD of the samples that did not support the deletion.

Deletions were then merged across all samples using 50% reciprocal overlaps and connected components. The merging process generated confidence intervals for the start and for the end position of each deletion that were used for further filtering, *i.e.,* if the upper confidence limit for the end position was lower than the lower confidence limit for the start position, or the outer confidence limits were smaller than the largest estimated deletion size in that region.

To filter the calls additionally after merging, a median threshold based on the sample libraries was computed for each sample, which in turn was used to calculate a combined threshold for each site depending on the samples that supported the deletion site. This combined site dependent threshold represents the minimum deletion size that is detectable by the samples that support this deletion. For convenience this threshold was used for further site filtering: (i) if the estimated deletion size was less than this threshold; (ii) if the discrepancy between the largest and the smallest deletion estimates from different samples was greater than twice the threshold; (iii) if the absolute difference between the deletion size estimate and the inner confidence interval was greater than

twice the threshold; (iv) if the inner confidence interval of the breakpoints was less than the threshold; (v) if the confidence limits of the deletion breakpoints were larger than twice the estimated deletion length. Additionally, sites were filtered out if the average RD ratio of the merged calls was greater than 0.75.

The confidence intervals for the breakpoint positions were determined by using the outer confidence limits as anchors. The uncertainty around those breakpoints was computed by using the discrepancy of the distance between the outer confidence limits and the deletion size estimates.

The false discovery rate (FDR) was estimated using SNP array probe intensities from the Omni 2.5 chip, which was run on samples from the 1000GP, together with the Genome STRiP Intensity rank sum (IRS) test[2]. To reduce the FDR of the callset, the results of the IRS test were used as training set. A likelihood ratio was computed for each deletion by fitting density curves for the attributes deletion size, BD score, number of supporting samples and estimated RD ratio. The curves were fitted separately for deletion calls that pass the IRS test and deletions calls that do not pass.

Deletion callset genotyping: Given the set of BreakDancer deletion sites initially inferred, we re-investigated all samples and independently generated the genotype likelihoods with congruent genotypes for each deletion in each sample. The genotype likelihoods were computed based on two major signals: 1) the number of discordant read pairs; and 2) the reduction of read-depth in the deletion compared to the flanking regions.

The discordant read pairs were identified by searching area left and right of the deletion breakpoints and seeking read pairs with insert sizes at the large extremes of the library's insert size distribution (z-score >3). If the SV size was small (<3000 bp), the searching area was chosen to be 500 bp long. Otherwise, the searching of the left end of a discordant pair was extended to 1500 bp upstream the first breakpoint in order to account for imprecision in breakpoints. We then summed for each deletion the number of read pairs sandwiching the deletion. That is, we counted inserts only if the right end of the discordant pair was within 500 bp downstream the second putative breakpoint.

To measure the read-depth reduction, we evenly sampled 20 positions in the deletion region and in each of the two flanking regions. The length of the flanking region was chosen to be the same as the length of the deletion unless the deletion was too close to the ends of the chromosome to accommodate such length. This allowed the comparison to be limited to a local context and robust to greater-scale copy number alterations. For smaller deletions (length <1000), we sampled fewer positions requiring that adjacent positions were 50 bp apart. This was to reduce the depth correlation between sampled positions. From the sampled positions, we compared read-depths computed in the deletion region and in each of the two flanking regions based on the Mann-Whitney U statistics. The likelihood contribution from read-depth reduction was based on the stratification of P-values of calculated statistics and median of read-depths in the three regions.

The integration of the two signals into genotype likelihoods comprised of three sources of evidence, 1) the relative scale of reduction in read-depth in the deletion; 2) the number of discordant read pairs; and 3) whether the reduction was statistically significant. Or formally,

$$\mathcal{L} = P(m_d|G)P\left(n_{drp}\middle|G\right)P(p|G),$$

where $G \in \{0,1,2\}$ corresponds to the three genotypes: (i) homozygous reference (no deletion), (ii) heterozygous variant, and (iii) homozygous variant. $P(m_d|G)$ denotes the probability of seeing the median of the read-depth in the deletion (m_d) given the median of read-depth in the flanking regions (m_f). This was modeled in a Gaussian density with mean at $\max(m_d, m_f), m_f, m_f/3$ and 0 in case $G = 0,1,2$ and standard deviation $m_f/2$. $P\left(n_{drp}\middle|G\right)$ denotes the probability of observing n_{drp} discordant read pairs given the genotype and the read-depths at the flanking region. The probability was calculated using a Gaussian density with mean at $m_f/3$ and standard deviation 1. The choice of $m_f/3$ instead of the ideal $m_f/2$ was to account for attrition of reads owning to the difficulty in mapping the discordant read pair and insert sampling in the deletion region. $P(p|G)$ denotes the probability of finding the reduction of read-depth in the deleted region statistically significant. $p = \max(p_1, p_2)$ where p_1 and p_2 denote the p-value of Mann-Whitney U test conducted between the deleted region and the two flanking regions. In other words, we required the reduction of the read-depth compared to both flanking regions to be indicative of true variants. This probability was obtained based on stratification described in the Table 2.1.1.

Based on the raw genotype likelihood (uniform prior), we generated genotype for all the deletions in all the samples. We assessed our genotypes in sites contributed to the phase 3 release against microarray intensities using the IRS Annotator implemented in GenomeSTRiP.

Table 2.1.1

P	P(p|G)
[0, 0.05)	1.0
[0.05, 0.1)	0.9
[0.1, 0.2)	0.1
[0.2,0.3)	0.01
[0.3, 0.5)	1e-4
[0.5, 0.8)	1e-8
[0.8,1.0]	1e-50

2.2. Delly

SV Type(s): DEL, DUP, INV

Contributed by: Tobias Rausch, Markus Fritz, Jan Korbel

Deletion callset generation: Delly[3] was run separately per population on all phase 3 low-coverage WGS samples. This tool uses paired-end mapping and split-read refinement to discover deletion sites in the genome. Delly first computes the insert size distribution of all input libraries and then uses an insert size cutoff of five times the median absolute deviation to classify deletion-supporting paired-ends. Paired-ends indicative of a deletion are then clustered together and refined using split-reads. All precise and imprecise Delly deletion predictions from the 26 populations were merged into a single structural variant site list using a 70% reciprocal overlap (RO) threshold and a maximum breakpoint offset of 250 bp. In each cluster, the paired-end call with the highest support was selected for the Delly's final candidate deletion site list.

Read-depth (RD) of all candidate deletions was annotated using 'cov', an auxiliary tool from the Delly package. The raw read-depth values were normalized for GC content, mappability and median total coverage across samples. For each candidate deletion, a Gaussian Mixture Model was applied to model the read-depth distribution and assign copy number states. Samples were genotyped using the posterior probabilities of the Gaussian Mixture Model, where samples with a posterior probability <0.9 were left ungenotyped. Filtering of candidate deletion sites was dependent on the quality of the Gaussian Mixture fit and the cluster separation. Using the copy number state assignments, a silhouette score was calculated and required to be >0.6 for a final deletion call. In addition, a minimum required ratio of genotyped compared to ungenotyped samples was set to 0.4 for each site. The read-depth modeling and filtering scripts used are both part of the current Delly distribution.

Tandem-duplication callset generation: Paired-end mapping and split-read refinement was used to discover tandem duplication sites in the genome with Delly. The tool was run separately by population on the phase 3 samples. Signal indicative of a tandem-duplication is a paired-end where the first and second read change their relative ordering compared to the expected Illumina paired-end library layout. These abnormal paired-ends are clustered together and refined using split-reads. All precise and imprecise Delly tandem duplication predictions were merged into a single SV sites list using a 70% RO threshold and a maximum breakpoint offset of 250 bp. In each cluster, the paired-end call with the highest support was selected for the final candidate tandem-duplication site list.

RD of all candidate duplications was annotated using 'cov'. The raw read-depth values were normalized for GC-content, mappability and median total coverage across samples. For each candidate tandem-duplication, a Gaussian Mixture Model was applied to model the read-depth distribution and assign copy number states. Samples were genotyped using the posterior probabilities of the Gaussian Mixture Model, where samples with a posterior probability <0.9 were left ungenotyped. Filtering of candidate tandem duplications was dependent on the quality of the Gaussian Mixture fit and the cluster separation. Using the copy number state assignments, a silhouette score was calculated and required to be >0.85 for making it into Delly's final tandem duplication sites list. Non-biallelic duplications with more than two read-depth components were filtered out as well as mixed sites.

Inversion callset generation: Inversions were identified by clustering all read-pairs of abnormal orientation compared to the standard Illumina paired-end layout. The left and right breakpoint of an inversion give rise to two different classes of inversion-supporting paired-ends that are clustered separately by Delly. Delly was used separately for each population of the 1000GP sample panel. Discovered population-specific inversions sites were subsequently integrated into a merged inversion site list using a strict 90% RO criterion and a breakpoint offset smaller than 50 bp. Left- and right- breakpoint spanning read pairs were initially merged independently (and, if feasible, subsequently joined into "two-sided inversions"; see below).

***Inversion callset genotyping**:* The merged inversion site list was genotyped across the entire 1000GP phase 3 cohort. Counts of inversion-supporting and reference-supporting read pairs were used to derive genotype likelihoods and phred-scaled genotype qualities.

Filtering of candidate inversion sites: All genotyped inversions were further filtered according to the below set of genotype quality metrics: (1) The minimum genotype ratio of genotyped to ungenotyped samples was greater or equal 0.4. (2) The fraction of inversion supporting pairs in carriers was greater or equal to 0.3. (3) The median carrier genotype quality (phred scaled) was ≥30. (4) The median non-carrier genotype quality (phred scaled) was ≥15. (5) All non-carriers were required to show zero inversion supporting paired-ends to filter inverted repeat induced false positive inversion calls. (5) The inversion size was greater than 250 bp and ≤50 kbp. Two-sided inversion sites exhibited confident support from both inversion breakpoints, and one-sided inversions showed support for one breakpoint only.

One-sided inversion sites were further filtered for split-read support by remapping reads around the predicted breakpoints (1 kb window) using bwa-mem[4]. For each site the median of the split read fraction across all carriers was determined. Using PacBio amplicon validation data an empirical fraction threshold was chosen (0.011), which minimized the FDR. Applying this threshold yielded Delly's final inversion call set.

2.3. VariationHunter

SV Type(s): DEL

Contributed by: Fereydoun Hormozdiari, Can Alkan, Evan Eichler

Deletion callset methods: VariationHunter[5] deletion discovery considered all discordant mapping locations (paired-end reads exhibiting mapping spans more than 4 standard deviations above the inferred mean insert size) from mrFAST and BWA read alignments. To generate an initial callset we considered only those candidate sites with support of at least two read pairs, whereby we required an average edit distance of maximum 3 per read. We then applied several filters to reduce false positives: 1) we scaled the minimum read pair threshold for each sample according to the depth of coverage; 2) removed deletion calls overlapping segmental duplications >30% (RO criterion); 3) removed deletion calls that also show inverted duplication or inverted repeat insertion signals; and 4) required the read-depth within the deletion interval to drop, consistent with the deletion event. Notwithstanding these filters, we considered a deletion call to be correct if it indicates an *AluY* or *L1HS* deletion, or it was also predicted as part of 1000 Genomes Phase 1 deletion callset. We further objectively filtered the callset based on total read support to reduce the FDR.

2.4. CNVnator

SV Type(s): DEL

Contributed By: Alexej Abyzov

CNV callset generation: SV calls with CNVnator[6] were made with standard parameters. Read-depth (RD) signals were corrected for GC bias. For each sample we aimed at using the smallest bin size out of the following three bin size values: 500, 1,000, or 1,500 bp, in such a way that the average RD would be at least 4 standard deviations away from zero. RD signal in 94 samples did not satisfy the criterion of 4 standard deviations for neither of the bins sizes, and thus these samples were not used for deletion discovery with CNVnator.

We then searched for read pairs showing abnormal read mapping with a mapping quality of at least 10, in support of the CNVnator calls, considering read pairs to be in support of a deletion if their mapping was consistent with a deletion and the span between reads showed an 80% RO with the CNVnator (read-depth based) SV call. Call bounds were readjusted to reflect the more precise breakpoint inference of paired-end mapping when compared to read-depth analysis.

For each sample we subsequently selected confident CNVnator calls as follows: 1) calls having paired-end support; 2) calls with p-values less than 10^{-5} (to accounts for multiple hypothesis testing, *i.e.*, calling in ~2,500 samples), and with $q0 < 0.5$; 3) deletion calls with p-values less than 10^{-5} and $rd*(1 + q0) < 0.75$ — whereby *rd* is the read-depth normalized to genome average, and *q0* is fraction of reads mapped with 0 (zero) mapping quality. We merged CNV calls for individuals within each population. For CNVnator site merging we initially clustered confident overlapping calls and averaged coordinates of each bound, pursuing the merging initially by population and then across the entire sample set.

2.5. Read-Depth (SSF)

SV Type(s): DEL, DUP, mCNV

Contributed by: Peter Sudmant, John Huddleston, Brad Nelson, Evan Eichler

CNV callset generation: The University of Washington (UW) read-depth based callset was generated subsequent to mapping all individual phase 3 short DNA-read genomic datasets with the *mrsFAST* read aligner[7] (using default parameters). Reads were first subdivided into their 36-bp non-overlapping constituents to normalize among the different read lengths represented in the 1000GP dataset. After mapping, read-depths were quantified for each genome and recalibrated to take into account GC-associated coverage biases introduced by library construction[8] and copy number was estimated in adjacent windows of 500 bp of unmasked sequence using a calibration curve based on regions of known copy number. Genomes were then assessed for overall quality using a number of QC metrics[9] with a total of 2169 samples passing all filters for analysis.

Calls were generated using digital comparative genomic hybridization (dCGH)[10] where the estimated copy numbers of each 'test' individual are compared against an ensemble of 'reference' individuals. In this case we used a set of 25 high-coverage, high-quality

reference genomes sequenced as part of the Denisova sequencing project[11]. The dCGH signal was segmented as described previously[10] using scale space filtering[12]. Briefly, the dCGH signal $f = \left(log_2 \frac{genome_{test}}{genome_{ref}}\right)$ was transformed to its Gaussian smoothed derivative $g'(\sigma)$ and second derivative $g''(\sigma)$ for a range of values of σ thus constructing a surface parameterized by the scale parameter σ. This surface is then traversed for local minima traversing from large σ to smaller σ thus indicating putative change-points of relative changes between the test and reference individuals. Change-points identified among each *test* (phase 3) sample and all 25 *reference* individuals were merged and a callset was generated for each individual genome. Finally, calls amongst all 2,169 *test* samples were simultaneously merged and genotyped to construct a callset. Genotypes were assigned using a Gaussian Mixture Model fit using expectation maximization. From the genotypes an assessment of the quality of each call was generated which we call the L-score, which is the sum of the log-probabilities of each samples genotype given the assumed genotype model.

The initially resulting UW read-depth sites list consisted of 24,655 sites overall, including 11,124 duplications, 7,019 deletions and 6,512 mCNVs. The FDR of these calls was estimated using the IRS method[13] and Affymetrix SNP chip data, and found to be higher than the initial inclusion threshold of 10%. We thus assessed the FDR for varying L-score cutoffs to generate the final SSF callset.

L-score cutoffs were set at 280, -180 and 880 for mCNVs, deletions and duplications respectively resulting in FDR <10%.

2.6. Genome STRiP

SV Type(s): DEL, DUP, mCNV
Contributed by: Bob Handsaker, Steve McCarroll

Deletion callset generation: The deletion discovery pipeline in Genome STRiP[14] version 1.04.1225 was used to discover and genotype large deletions in 2535 samples sequenced at low coverage (these initially included 31 samples from the 1000GP not used in the final phase 3 release, since they exhibited patterns of cryptic relatedness).

Deletion discovery was performed grouping the samples in five separate batches of 500 samples each (the last batch had 535 samples), with a target deletion size range of 100 bp to 100 kbp. After discovery, the union of the discovered deletion sites was genotyped in all samples simultaneously. Standard Genome STRiP genotyping filters were applied to select passing sites and to remove duplicate calls and then a more stringent duplicate removal protocol was applied.

For deletions larger than 100 kb, we applied the same method described above using a target deletion size range of 100 kb to 1 Mb. For these larger sites, we used a more

stringent site selection threshold post-genotyping, requiring a read-depth cluster separation >6 standard deviation. In addition, three large deletion sites >100 kb that appeared to overlap a true deletion but had incorrect boundaries were removed during manual review.

Genome STRiP employs a genotype-likelihood-based method for detection and removal of duplicate calls. All pairs of calls are evaluated based on the degree of overlap and the degree of genotype concordance as represented by the genotype likelihoods. When calls are deemed to be duplicates, the preferred call is chosen to maximize the posterior probabilities of the genotypes.

The default protocol for duplicate call detection requires site overlap greater than 50% and duplicate score (logarithm of odds (LOD) score of genotype concordance at most discordant sample) greater than zero. This is a conservative threshold intended for input site lists with a low level of duplicate calls. In conjunction with performing deletion discovery in batches, we applied a more stringent protocol for duplicate detection:

a) Remove all duplicate calls using standard settings (50% site overlap and most-discordant LOD score greater than zero).
b) Perform a second pass removing duplicate calls using criteria of site overlap greater than 50% and no discordant genotypes at a 95% confidence threshold.
c) Perform a third pass removing duplicate calls using criteria of 80% site overlap only.

In addition to the deletions ascertained and genotyped using the Genome STRiP deletion discovery pipeline, we included in the deletion discovery set bi-allelic deletion sites on the autosome ascertained through the Genome STRiP copy number variant (CNV) discovery pipeline when these deletion sites had less than 50% overlap with deletions already ascertained by Genome STRiP.

mCNV callset generation: A pre-release version of Genome STRiP CNV discovery pipeline (version 1.04.1375) was used to discover and genotype large copy number polymorphisms, including deletions, duplications and mixed deletions/duplications. The Genome STRiP CNV discovery pipeline utilizes primarily read-depth during discovery and is complementary to the Genome STRiP deletion discovery pipeline.

The reference genome was divided into 1,061,745 overlapping windows, each consisting of 5 kb of uniquely alignable sequence and overlapping adjacent windows by 2.5 kb. These windows were genotyped using Genome STRiP and windows with evidence of polymorphism were retained. Adjacent or overlapping windows with compatible genotypes were merged to increase power. After this genome-wide scan, 179 samples exhibiting excessive variation were removed from this analysis based on the number of distinct calls in these samples exceeding the median (across all samples) + 3 MAD. The effective discovery cohort contained 2,356 samples.
For each window with evidence of polymorphism, a hill-climbing algorithm was used to refine the variant boundary through multiple rounds of genotyping using different boundary intervals. The objective function used to select the best interval maximizes the

sum of genotype confidence for samples with non-modal copy number for the site. After boundary refinement, duplicate calls are removed using the standard Genome STRiP duplicate removal settings (50% site overlap and discordant LOD score of zero).

Candidate calls were then filtered using the following criteria: Sites are retained if (a) they had a 95% confident genotype call rate of at least 80% (b) they had at least one sample called non-homozygous-reference at 95% confidence, (c) at least 30% of the covered bases were uniquely alignable and (d) the read-depth cluster separation was at least 5 standard deviations. Sites passing these filters that were larger than 10 kb were retained. Sites between 3 kb and 10 kb were retained if IRS p-values could be computed and all available p-values were less than 0.01.

For the sites that were not confidently called as bi-allelic deletions, copy number likelihoods were converted to multi-allelic genotype likelihoods. An expectation-maximization algorithm was used to estimate the haploid allele frequency of each copy number state, assuming Hardy–Weinberg equilibrium (HWE) within each population. Alleles with an overall posterior likelihood of at least 0.001 were used in the variant model and genotype likelihoods for this set of alleles were generated based on the diploid copy number likelihoods and the estimated allele frequencies in each population.

Calling of SVs on chromosome X was similar to the procedure described above for the autosome, except that processing was carried out separately after grouping samples into batches to control for differential read-depth profiles on chromosome X that were primarily driven by differences in sequencing depth between blood-derived DNA samples and DNA samples from lymphoblastoid cell lines.

CNV discovery and genotyping was performed based on dividing the samples into four batches: F1 and F2 for female samples with normalized chrX dosage below/above 1.96 (respectively) and M1 and M2 for male samples with normalized chrX dosage below/above 0.9985. Discovery was performed separately in the F1, F2 and M1+M2 cohorts. Following the QC procedures outlines above for each of the three discovery batches, samples were dropped if the number of called sites were above the median + 3 MAD in any of the three batches or in the autosome. This yielded a discovery cohort for chrX of 2,137 samples (1,054 males and 1,083 females).

The union of the discovery sites was then genotyped in two batches, one batch containing the F1+M1 samples and one batch containing the F2+M2 samples. The genotyped sites were merged and and duplicate site removal was performed as described above. Sites larger than 20 kb were retained based on an estimated FDR using the IRS method of 2.5% (Omni 2.5 array) and 0% (Affy6 array). Sites between 3 kb and 20 kb were retained if IRS p-values could be computed and all available p-values were less than 0.01. The final callset on chrX consisted of 764 sites of which 392 were confidently classified as bi-allelic deletions.

A total of 74,751 potentially redundant input sites (autosome + chrX) were merged and genotyped using Genome STRiP version 1.04.1257 to generate a set of 32,924 mostly non-

redundant sites with genotype likelihoods in 2,535 samples suitable for building the imputation scaffold with Impute2.

2.7. Pindel

SV Type(s): DEL

Deletion callset generation (contributed by: Kai Ye, Eric Wubbo Lameijer, Klaudia Walter): Pindel[15] (version 0.2.5a2) was run across Illumina paired-end samples in chunks of 300 kb with the following parameters: -w 0.1 -x 5 -B 0 -T 4. Regions around the centromeres were excluded. Split read based deletion calls appearing in at least five samples and with more than five reads from both strands were collected for downstream analysis. We estimated the FDR for deletions greater than 300 bp using SNP array probe intensities from the Omni 2.5 and the Affymetrix 6.0 chips (both run on samples from the 1000GP) together with the Genome STRiP IRS test. To reduce the FDR of the deletion call set, the results of the IRS tests were used as training set. A likelihood ratio was computed for each deletion by fitting density curves for the attributes deletion size, length of the micro-homologies around the breakpoints, number of supporting samples and percentage of P-sites (the sites that passed all filters from the strict genome mask annotation). The curves were fitted separately for deletion calls that passed the IRS test and deletions calls that did not pass. 14 deletions were randomly selected for PCR and Sanger sequencing. We did not observe a PCR product for one deletion, and 11 out of the remaining 13 deletions showed the (exact) same breakpoints as predicted by Pindel.

Complex deletion callset generation (contributed by Kai Ye, Ali Bashir): In addition to the aforementioned Pindel deletion callset, we employed the latest Pindel version also for capturing deletions with inserted sequences at the deletion breakpoints (*i.e.*, complex deletion events) establishing an additional callset with Pindel complex deletion sites not included in our phase 3 SV group data release (complex Pindel deletions available at ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map). Complex Pindel deletion sites were called across the entire set of low-coverage Illumina samples, and variants retained within the complex Pindel callset if we observed read support evidence from both strands and further if (and only if) the variant was present in NA12878.

2.8. MELT

SV Type(s): MEI
Contributed by: Eugene J. Gardner, Scott E. Devine

MEI callset generation: Mobile element insertions (MEIs) were detected with the Mobile Element Locator Tool (MELT)[16] using discordant read pairs (DPs) to define potential MEI sites and split reads (SRs) to identify breakpoints and target site duplications (TSDs). MEIs

were detected across all phase3 Illumina low-coverage sample binary alignment/map (bam) files. Samples with less than 90% properly mapped read pairs were removed from the analysis because high levels of mapping artifacts in these samples confounded MEI detection. The 82 samples that were excluded are listed in Table 2.8.1 below. Imputed MEI calls were included in the final SV call set for the 82 samples in Table 2.8.1. MELT can be obtained from http://melt.igs.umaryland.edu/.

MELT pipeline steps*:* DPs were first extracted from bam files and subsequently screened to identify mate pairs where one mate mapped unambiguously to the reference human genome sequence (Reference Mate – RM) and the remaining mate (Mobile Mate – MM) aligned to one of three mobile element reference sequences (282 bp *Alu* Y consensus[16], 6,019 bp L1-Ta L1.3[17], or 1,628 SVA[18]) using Bowtie2[19] with default parameters. MMs were further refined after alignment by developing filtering cutoffs to accommodate normal sequence variation in each mobile element type[16]. Sites where at least four RMs clustered within 500 bp of each other were considered candidate MEI sites. DPs and SRs that mapped to each candidate site then were merged across all samples to build models containing all available evidence for each candidate MEI site. These models were used to identify the following features at each MEI site: the precise insertion site, strand, TSD, insertion sequence and insertion length. All putative MEIs were genotyped across the 1000GP Phase 3 samples using a modified version of the equation described in Li *et al.*[20].

Filtering of Candidate MEIs: MEIs were required to have at least four DPs of supporting evidence during initial discovery at each site for the final call set. This provided a good balance between the false negative rate (FNR) and the FDR (Table 1, main text). Putative MEIs were filtered if they mapped within reference mobile elements of the same type as annotated by RepeatMasker v. 4.0.3 at the University of California Santa Cruz (UCSC) Genome Browser website[21]. To control for sequence coverage variation at candidate MEI sites, 100 bp windows flanking each MEI site were sampled for depth of coverage fluctuations. Sites that fell outside of the range of 70 to 130% sequence coverage were filtered.

Table 2.8.1 - MELT Filtered Samples

GENOME NAME	PERCENT READS PROPER PAIRED	GENOME NAME	PERCENT READS PROPER PAIRED
HG01182	35.65383262	NA19914	84.10528923
HG01183	37.74160819	NA18636	84.43989059
HG01187	40.5099555	NA19473	84.64369818
HG01188	48.38791739	HG00236	84.80342026
NA19474	50.08947675	HG00140	84.89970785
NA19338	54.74811848	HG00238	85.0720453
NA19055	56.49774621	NA19430	85.13356496
NA18498	59.51585106	NA19247	85.49534598
HG01522	60.52264324	NA19248	85.80111166
NA19707	63.54952663	NA19901	86.14479404
HG01495	63.717762	NA18870	86.19125899

HG01204	64.35374775	HG01366	86.24589902
NA19058	64.45264614	HG01334	86.49511825
NA18867	65.53874032	NA18853	86.79407677
NA19116	67.5861823	HG02121	86.90209268
NA18984	69.10197844	NA19064	87.32160657
NA19000	73.44250778	HG01437	87.68970225
HG00123	73.45059672	NA06985	88.13133853
NA18982	74.26809319	HG03907	88.4480774
HG01101	74.80170209	NA19819	88.6504557
NA18537	75.9427905	HG03925	88.652049
NA18986	76.4162398	NA19075	88.69913482
NA11994	77.02959592	NA19213	88.77173126
HG00110	78.46756493	HG03595	88.9682155
NA18504	78.58029335	HG03908	88.99989296
NA19719	79.22589977	HG03805	89.23752353
NA18624	79.82318449	NA19682	89.27943855
NA18912	80.16712307	HG03594	89.30545592
NA18632	80.89482933	HG03926	89.384562
HG01168	81.80151954	NA19189	89.40531169
NA19063	81.92548278	HG02008	89.41626207
NA19703	82.83596144	HG01176	89.44784702
NA19789	82.9684615	NA19060	89.55853257
NA18633	83.00193183	HG03304	89.56506047
NA19909	83.03416108	HG01167	89.56741161
NA18623	83.27101572	HG03378	89.64475573
HG00864	83.39286482	NA19009	89.65493494
NA12399	83.4811382	HG03354	89.67373574
NA20800	83.60649987	NA19066	89.77061762
NA12400	83.73956843	NA19921	89.8945942
NA11832	83.9340582	NA19236	89.95338935

2.9. Dinumt

SV Type(s): NUMTs

Contributed by: Gargi Dayama, Ryan Mills

NUMTs callset generation: Nuclear insertions of mitochondrial DNA (NUMTs) were discovered using dinumt[22] (version 0.0.22) in 1000GP phase 3 samples using the following parameters: --len_cluster_include = mean + 3 * standard deviation of sample insert size, --len_cluster_link = 2 * len_cluster_include, --max_read_cov = 5 * mean sample coverage. When possible, soft clipped reads were used to identify breakpoint positions. Confidence intervals were set to the distance between most prevalent clipped positions, if available;

otherwise, confidence intervals were set to the inner distance between supporting read pair clusters. Calls were then filtered based on the following criteria: phred-scaled quality filter <50, number of supporting reads <4, manual inspection.

NUMTs callset genotyping: Filtered calls were merged across samples and genotyped in the same samples using custom software (gnomit, version 0.0.22) using a similar approach to previously applied methodology[20] to calculate genotype likelihoods, with breakpoint position refinement based on cross-sample support and mtDNA mapping of soft clipped reads. Population frequency was estimated from initial likelihoods and used as priors in subsequent expectation maximization iterations (max 10). No-calls or calls with genotype quality <13 were labeled as LowQual. Calls labeled with "IMPRECISE" did not have a refined breakpoint position. 80 samples were omitted from genotyping due to sequencing quality – NUMT calls in these have been included based on imputation at the stage of haplotype phasing. Dinumt is available at: https://bitbucket.org/remills/dinumt.

2.10. Summary of all callsets

Contributed by: Tobias Rausch and Peter Sudmant

The total number of calls made by each method is presented in Table 2.10.1 along the diagonal with the number of common calls by any two methods on the off-diagonals. We note that that due to our strict FDR cutoffs many call-sets were heavily pre-filtered (e.g., for Pindel only SVs < 1 kbp). So the true overlap among the *raw* calls is much higher than what is shown in this table. Callers pursuing population-based SV discovery are indicated with an asterisk (*). All the other five SV detection algorithms made calls independently per sample (*i.e.,* did not perform population-based calling).

Table 2.10.1: SV calling algorithms, the total number of calls made by each tool, and calls common to pairs of callers.

Methods	Break-Dancer*	CNV-nator*	Delly*	GenomeSTRiP	MELT	Dinumt	Pin-del	SSF	Variation-Hunter*
BreakDancer (DEL)	10552	4925	3029	9738	0	0	150	1866	7565
CNVnator (DEL)	-	18345	5056	12086	0	0	9	6800	11133
Delly (DEL, DUP, INV)	-	-	8229	6948	0	0	28	364	6222
GenomeSTRiP (DEL, DUP, mCNV)	-	-	-	38404	0	0	417	1262	16042
MELT (*Alu*, L1, SVA)	-	-	-	-	16631	0	0	0	0
Dinumt (NUMTS)	-	-	-	-	-	168	0	0	0
Pindel (DEL)	-	-	-	-	-	-	9580	0	276
SSF (DEL, DUP, mCNV)	-	-	-	-	-	-	-	4082	367
VariationHunter (DEL)	-	-	-	-	-	-	-	-	23528

We additionally prepared a release of our SV map lifted over to GRCh38 coordinates. All but 98 SVs (0.1%) could be lifted over, with only 13 SVs being lifted to an unplaced GRCh38 contig, only 10 SVs differing in size by more than 10%, and the breakpoints of 99.7% of assembled deletions were identical between GRCh37 and GRCh38. At most 0.5% of SVs in our GRCh37 release have been fully or partially incorporated into GRCh38.

3. SV site merging, genotyping and phasing

Contributed by: Bob Handsaker, Tobias Rausch

To generate a haplotype-resolved SV set we used the following procedure. Initially, in order to generate a high confidence set of large deletion sites to be used for joint haplotype scaffold generation along with SNPs and insertions/deletions (indels), we employed Genome STRiP[14] to re-genotyping sites called with the five most specific deletion discovery algorithms (BreakDancer[1], Delly[3], CNVnator[6], GenomeSTRiP[14], and VariationHunter[5]). GenomeSTRiP's redundancy removal function was used to merge these sites into a coherent list of large high confidence deletions.

A total of 74,751 potentially redundant input sites (autosome + chrX) were merged and genotyped Genome STRiP version 1.04.1257 to generate a set of 32,924 mostly non-redundant sites with genotype likelihoods in 2,535 samples suitable for building the imputation scaffold with Impute2.

As the input call sets originated from multiple algorithms using the same input data, so a large degree of redundancy was expected. In addition, the CNVnator input sites were the union of sites called separately in each population and were therefore expected to also have a high rate of internal redundancy.

To resolve this high rate of redundancy, we used a stringent protocol for duplicate site detection and removal:

a) Remove all duplicate calls using standard settings in Genome STRiP (50% site overlap and most discordant LOD score greater than zero).
b) Perform a second pass removing duplicate calls using criteria of site overlap greater than 50% and no discordant genotypes at a 95% confidence threshold.
c) Perform a third pass removing duplicate calls using criteria of 80% site overlap only.

The results of each merging pass are indicated in the three summary lines in Table 3.1 (Merge1, Merge2, Merged call set). FDR estimates (using the IRS method, Omni 2.5 array) ranged from 1% to 6.7% in the input call sets. The estimated FDR in the genotyped call set was 3.1%.

Table 3 .1– Merged Deletion Genotype Statistics

Call set	Input FDR Estimate	Input Sites	Dup. Sites	Mono Sites	Other Filters	Pass Sites	Pass %	IRS Sites Evaluated	IRS Eval %	Pass Site IRS FDR Estimate	Self-overlap @ 50%	Self-overlap @ 80%
EM_DL	4.4%	7,099	0	118	127	6,854	98.2%	4,750	69.3%	1.3%	0.0%	0.0%
SI_BD	3.2%	17,865	39	211	6,964	10,651	60.5%	6,226	58.5%	2.1%	0.0%	0.0%
YL_CN *	1% - 5%	41,294	27,269	53	1,128	12,844	91.9%	8,199	63.8%	1.6%	13.7%	11.7%
UW_VH2	6.7%	20,254	2,794	185	2,817	14,458	83.7%	9,758	67.5%	2.5%	1.7%	1.0%
BI_GS	2.5%	29,944	0	0	0	29,944	100.0%	18,376	61.4%	2.5%	1.2%	0.0%
Merged 1	-	74,751	36,547	0	0	38,204	100.0%	21,401	56.0%	3.0%	14.9%	12.9%
Merged 2	-	74,751	40,955	0	0	33,796	100.0%	19,670	58.2%	3.2%	4.1%	2.5%
Merged call set	-	74,751	41,827	0	0	32,924	100.0%	19,567	59.4%	3.1%	1.6%	0.0%

* The CNVnator input calls were the union of 26 callsets made separately in each population.

The merged SV list was used for haplotype scaffold generation, along with SNPs and bi-allelic indels, using ShapeIt2[23]. Following scaffold generation, all other SV callsets were statistically phased into these haplotype scaffolds using MVNcaller[24], with the confidence in individual variant phases depending on patterns of LD and VAF (e.g. singletons are arbitrarily phased in this procedure). 98% of SVs >1% have a reported median MVNcall SV carrier (phased genotype) posterior probability[24] >0.95, compared to 89% for variants <1%. Finally, we performed another SV merging and filtering step in order to remove redundant calls, to harmonize the SV notation and to ensure a low site for the merged SV call set. All post-phasing mono-monomorphic reference sites were excluded, cryptically related samples (*N*=31) were dropped and CNVs were classified as bi-allelic deletions (DEL), bi-allelic duplications (DUP) and multi-allelic CNVs (mCNV). Merging was performed using an overlap graph G(r, c) = G(0.71, 0.71), requiring a RO (r) of at least 71% and a non-reference copy number concordance (c) of at least 71%. Using these cutoffs ensured that >99% of all connected components in the overlap graph were cliques. For each connected component, we picked one representative call whereas all merged calls were specified in the VCF INFO column.

After callset merging and statistical haplotype phasing we obtained improved site-based FDR estimates (these are summarized in Table 1), which likely is due to a reduction of samples with missing genotypes and improved SV boundary inference after merging.

4. Breakpoint Analysis

4.1. Local assembly (short-read)

TIGRA (contributed by: Ken Chen, Wanding Zhou, Zechen Chong, Xian Fan)*:* Breakpoint assembly of deletions and duplications in our callset was performed using TIGRA-0.4.0[25]. For each breakpoint, TIGRA first extracted reads that were mapped near the predicted breakpoint (±500 bp) from the set of bam files corresponding to carrier samples. Paired-ends that were unmapped or mapped outside the window were also extracted. We then ran an iterative *de Bruijn* graphic assembly algorithm to decode the set of non-reference alleles that best explain the set of reads. An assembly score was calculated to summarize both the length of the contigs and the proportion of reads that contributed to the assembly.

Velvet (contributed by Amina Noor, Danny Antaki, Madhusudan Gujral, Jonathan Sebat): To enable further characterization of SV complexity, DEL and DUP calls were generated in PCR-free high coverage Illumina WGS data for 30 samples using the forestSV[26] tool. A total of 1,248 non-redundant raw DEL and DUP calls were initially made for this purpose (including simple and complex sites). Genotype likelihoods for these calls were determined using expectation maximization Gaussian Mixture Model classifier. The calls were then filtered with non-reference genotype likelihood of greater than 0.75, resulting in 1,148 deletion and 43 duplication calls. FDR was determined to be 5.4% based on the IRS test on array intensity data. To assemble breakpoints, soft-clipped reads were extracted within +/- 1 kbp of start and end positions identified by forestSV. De novo assembly of breakpoint-spanning reads was performed using Velvet[27]. Assembled contigs were aligned to the reference genome using BLAT[28] and breakpoints were inferred from the alignments.

For our call set, 4,838 putative breakpoint contigs were assembled and BLAT alignments identified non-redundant breakpoints for 419 calls (indicating a success rate of 36%). We verified breakpoints for NA12878 using Moleculo data, and observed that 40 out of 40 breakpoint contigs were supported by an identical Moleculo long-read.

SVs having multiple breakpoints were identified as complex breakpoints and there were 69 such instances. We classified these breakpoints by performing BLAT onto the reference genome and generating dotplots for each SV. Out of these, 3 were deletion with inversions, 22 were insertions within deleted sequence, and 44 were multideletions.

4.2. Long read based breakpoint analysis

PacBio SMRT sequencing of NA12878 (contributed by: Ali Bashir, Matthew Pendleton, Robert Sebra, Gintaras Deikus, Eric Schadt, Chris Mason): As described in further detail elsewhere[29], aliquots of 5 µg of NA12878 genomic DNA (Coriell) were diluted to 150 µL using Qiagen elution buffer at 33 µg / µL. The 150 µL aliquot was individually pipetted into the top chamber of a Covaris G-tube spin column and sheared gently for 60 seconds at

4,500 rpm using an Eppendorf 5424 bench top centrifuge. Once completed, the spin column was flipped after verifying that all DNA was now in the lower chamber. Then, the column was spun for another 60 seconds at 4,500 rpm to further shear the DNA and place the aliquot back into the upper chamber, resulting in a 10,000 to 20,000 bp DNA shear, verified using a DNA 12000 Agilent Bioanalyzer gel chip. The sheared DNA was then re-purified using a 0.45X AMPure XP purification step (0.45X AMPure beads added, by volume, to each DNA sample dissolved in 200 µL elution buffer (EB), vortexed for 10 minutes at 2,000 rpm, followed by two washes with 70% alcohol and finally diluted in EB). This AMPure XP purification step assures removal of any small fragment and/or biological contaminant.

After purification and shearing, ~1.6 to 3.2 ug of purified and sheared sample was taken into DNA damage and end-repair from each batch preparation. The DNA fragments were repaired using DNA damage repair solution (1X DNA damage repair buffer, 1X NAD+, 1 mM ATP high, 0.1 mM dNTP, and 1X DNA damage repair mix) with a volume of 21.1 µL and incubated at 37°C for 20 minutes. DNA ends were repaired next by adding 1X end repair mix to the solution, which was incubated at 25°C for 5 minutes, followed by the second 0.45X Ampure XP purification step. Next, 0.75 µM of blunt adapter was added to the DNA, followed by 1X template preparation buffer, 0.05 mM ATP low and 0.75 U/µL T4 ligase to ligate (final volume of 47.5 µL) the SMRTbell adapters to the DNA fragments. This solution was incubated at 25°C overnight, followed by a 65°C 10-minute ligase denaturation step. After ligation, the library was treated with an exonuclease cocktail to remove un-ligated DNA fragments using a solution of 1.81 U/µL Exo III 18 and 0.18 U/µL Exo VII, then incubated at 37°C for 1 hour. Two additional 0.45X Ampure XP purifications steps were performed to remove <2000 bp molecular weight DNA and organic contaminant. The exonuclease cycle above was repeated a second time on all library preparations and followed by an additional two 0.45X Ampure XP purifications and a third 0.40X Ampure XP purification step, to chemically size select as stringently as possible.

Verification of MEIs, NUMTs and Complex Pindel calls using PacBio reads in NA12878 (contributed by: Ali Bashir): All phase 3 MEI and NUMTs calls in NA12878 were considered analyzed with PacBio data as another means of verifying the respective callsets. To process candidate complex Pindel sites, calls were first filtered using a minimum size (50 bp) and a minimum Levenshtein distance of 50 relative to the reference hg19 allele.

All raw NA12878 PacBio reads were aligned to hg19 using BLASR[30]. For each candidate SV, reads overlapping the region of interest were extracted. For complex Pindel calls, a synthetic reference was created representing the putative call at each locus. Reads were remapped to both the true and synthetic references; all reads that preferentially mapped to the synthetic reference were passed on to the next step of the pipeline (for larger MEI and NUMTs events this step was not required as the error-correction and assembly process would automatically separate more highly diverged alleles).

Filtered reads were passed through an error-correction process (as described below) to reconstruct putative assemblies spanning through each SV. Error-correction of all reads was performed following the general principles proposed in Chin *et al.*[31], using the FALCON pipeline (https://github.com/PacificBiosciences/FALCON; [2015]). In short, all long reads

are aligned to one another using BLASR. These reads are then grouped together by selecting the top alignments (using a coverage cutoff of 40). A consensus is formed for each read; the resulting read is trimmed at the ends to eliminate potential chimeras and low-quality sequence (here we required at least 5X coverage of a given base). The reads are then assembled using a string graph based assembly approach. Spanning assemblies were returned for 82% (893 of 1,094) MEI predictions, 100% (4 of 4) NUMTs predictions and 100% of complex pindel events (743 of 743).

The resulting assemblies are then separately aligned, using BLASR, to the sequence immediately left and right of the putative breakpoint in hg19 of the putative in hg19. Breakpoint precision (Table 1) was evaluated by comparing the observed junctions vs. aligned end points on each side of the breakpoint. For Pindel SVs, the putative insert sequence between breakpoints was additionally extracted. These insert sequences were then compared to the predicted Pindel calls using the Needle pairwise alignment tool from the EMBOSS package[32] in order to assess SV call accuracy.

4.3. Breakpoint derivation and analysis

We attempted de novo assembly of all deletions and duplications except for the Pindel calls that are based on split read analysis, and which thus were already reported at basepair-resolution in the original SV input callset. We evaluated the accuracy of Pindel's SV breakpoint assignments by comparing a subset of the SVs reported by Pindel with assembled breakpoints. This analysis revealed that Pindel breakpoints are in perfect agreement with our assemblies in >95% of cases (*i.e.,* for 2,646 out of 2,764 breakpoints of 1,382 overlapping calls). All assembled breakpoint and Pindel calls were merged and used for fine-resolution bp analysis.

CROSSMATCH alignments (contributed by: Ken Chen, Wanding Zhou, Zechen Chong, Xian Fan): Breakpoint assembly contigs were first aligned using CROSSMATCH (Green, unpublished; http://www.phrap.org/phredphrapconsed.html; [2015]) against the corresponding reference assembly region that spans the putative breakpoints with 700 bp flanking sequence on either end. A breakpoint was called "validated" and passed to the next stage if the associated pair-wise alignments indicated the existence of the same SV class (*e.g.*, deletion) as was predicted by the original callers. A breakpoint was not validated if the alignment was ambiguous, *i.e.*, containing more than two high scoring pairs and having an assembly score <200, or if the size of SV differed by more than 50% from the original prediction.

AGE alignments (contributed by: Alexej Abyzov, Daniel Rhee Kim, Ken Chen): Contigs locally assembled with TIGRA-SV[25] were aligned with AGE[33] to target deletion regions extended by 1 kbp downstream and upstream. AGE was run with options '-indel –both' to infer deletion breakpoints and with options '-tdup -both' to infer breakpoints of tandem duplications. The following scoring parameters were utilized: match=1, mismatch=-10, gap_open=-10, and gap_extend=-1. Typically each region had few alternative contigs assembled, and each one was aligned to the target region.

Each AGE alignment consists of aligned sequence from the left flank, an excised region (the candidate SV region), and aligned sequence from the right flank. For each deletion breakpoints were assigned as coordinates of excised region from a contig alignment that satisfies all the following requirements: (i) the contig is at least 100 bp in length; (ii) at least 90% of contig bases are aligned; (iii) length of each alignment flank is at least 35 bp (regions of sequence micro-identity around excised region are not included in the lengths calculation); (iv) contigs have no more than one alternative alignment of equal score; (v) average alignment sequence identity in right and left flanks should be at least 96%; (vi) alignment sequence identity in each flank should be at least 95%; (vii) coordinates of excised region and target region should overlap reciprocally by at least 50%; (viii) each coordinate (*i.e.*, start and end) of target region and excised region should not differ by more than 500 bp; (ix) for an alternative alignment the previous two requirements should also be satisfied. In case more than one contig satisfies the requirement no breakpoints were assigned to the SV.

Unified set of breakpoints: Breakpoints derived from CROSSMATCH and AGE alignments were merged and only SVs showing the exact same breakpoint junctions by both aligners were retained.

Validation of breakpoint by PCR: We performed two rounds of breakpoint validation with PCR. First we selected 40 random sites of deletions and designed primers to amplify breakpoint sequence. In three cases the primers did not result in amplicon bands, neither for the reference nor the alternate allele, and in three cases unspecific amplicons were generated. For the remaining 34 cases we sequenced the amplified band with Sanger technology, and in all cases the obtained sequence matched the one inferred through assembly. Second we investigated 72 deletions exhibiting an additional insertion ('micro-insertion') longer than 100 bases through PCR-amplification. In two cases no PCR product or an unspecific PCR product was yielded, in two cases Sanger sequencing of the amplified band failed, and in three additional cases we could not complete capillary sequencing through the entire micro-insertion sequence. Of the remaining 58 sites, one was invalidated, the sequence for 53 showed an exact match to the one predicted from assembly, and for 4 sites there was a single basepair or short indel difference in sequence (these differences could be resulting from polymorphisms existing in population).

Verification of breakpoints in high coverage genomes: For verification purposes 30 genomes were sequenced with PCR-free library preparation and 250 bp reads up to depth of 60X. We utilized this data for deletion breakpoint validation. For each deletion with assembled breakpoints and using the corresponding genotype information we selected high coverage samples carrying the deletion. For those samples read pairs with mapped coordinates in the 2 kbp vicinity of the breakpoints were extracted from BAM files, and each was tested for an overlap at the leftmost flanking ends. If a suitable overlap was detected, the reads were merged into a long continuous (gapless) genomic fragment of 250 to 450 bp in length. To avoid confounding factors affecting this validation exercise (like mis-genotyping and low efficiency of finding read overlaps) we only considered deletion sites genotyped in at least four of those 30 high coverage samples. Using AGE we realigned constructed long

fragments around breakpoints. We considered breakpoint perfectly confirmed if majority of the aligned fragments had exact match to breakpoint sequence. Based on analysis of 879 deletion sites, 94.9% of breakpoint sequences were reproduced exactly, 2.7% of breakpoints showed minor sequence differences, while 2.4% of breakpoint were invalidated (approximately half of these showed MIs longer than 100 bp – such MI is another confounding factor for this validation, as their length is comparable to constructed long fragments).

Breakpoint complexity analysis (contributed by: Ali Bashir and Alexej Abyzov):

To determine micro-insertion placement, 10 kbp upstream and downstream of the predicted phase 3 breakpoint junctions (including the original deletion interval) were analyzed for all assembled insertion sequences at least 10bp in length. The assembled insertion sequence was set as the reference and the hg19 interval as the query; Nucmer[34] was run on the two sequences using the parameters "nucmer -mumref -l 10 -c 10". The resulting alignments were then filtered using the delta-filter command: "delta-filter -1 -i 95" to determine optimal hits between query and the reference, ignoring hits with less than 95% alignment identity. We iterated through all remaining alignments and distinguished the following alignment categories: (i) The alignment is in an inverted orientation and abuts the deletion start (or end) boundary within 5 bp of the original deletion breakpoint; (ii) The alignment starts within the breakpoint interval or <3 bp before the beginning of the breakpoint interval, and ends within the breakpoint interval or <3 bp past the end of the alignment; (iii) The alignment starts downstream of the region or overlaps by <3 bp, or the alignment ends upstream of the breakpoint (or overlaps by <3 bp); (iv) The alignment overlaps either end of the original deletion boundary (the alignments were required to overlap >3 bp on each side of the breakpoint boundary to reduce spurious alignment calls). For all unaligned subintervals >= 22 bp, we realigned these substrings to hg19 using bwa-mem[4] in order to determine a potential origin for the sequence (the quality of each base was set to be q20). These mapping assignments were then used to assign categories to each event, as follows: *Ins and Del* – if the inserted sequence cannot be aligned to any interval, *Ins with Dup and Del* – the inserted sequence contains a single duplication from outside the original deletion interval; *Ins with MultiDup and Del* – the inserted sequence contains at least 2 distinct duplications outside the deletion interval; *Inv and Del* – if the inserted sequence is an inverted subsequence from the original deletion interval that is located at either boundary of the deletion; *MultiDel with inverted or non-inverted spacer* – if the inserted sequence maps to a single (or multiple) subintervals from the original deletion interval; *other* – it did not fit *strictly* into one of these categories.

Additionally, to determine NA12878 specific insertion sequences within deletions the same PacBio de novo assembly procedure was performed on all NA12878 phase 3 deletion breakpoints as as described previously (see "*Verification of MEIs, NUMTs and Complex Pindel calls using PacBio reads in NA12878"*). Given the short size cutoff for examining insertions (>=10 bp) an additional filter was added which required that the PacBio assembly was able to verify the original deletion boundary within 1 bp. This limited the total number of deletion breakpoints to 750.

Identification of SVs with complex breakpoints by joining adjacent SV calls (contributed by: Alexej Abyzov, Taejeong Bae): We reasoned that in some cases independently called SVs that are adjacent in the genome may actually correspond to a single complex structural variation event, and therefore can be reconstructed by joining multiple (overlapping and non-overlapping) adjacent candidate SV sites. To identify examples of breakpoint complexity in such manner, we extracted SV pairs with the following characteristics: the distance between them is less than 100 kbp, they are genotyped in at least five individuals, and the *r* value LD for them in the population is greater than 0.9. This analysis yielded 119 SV pairs, of which 31 were genotyped in NA12878.

Using long (2-10 kbp) PacBio and Moleculo reads for NA12878, we attempted to confirm that the candidate SV pairs are indeed observed in the same haplotype. We also attempted to derive their exact breakpoints with the associated complexity. We first re-estimated the boundaries of candidate pairs, based on visual analysis of read-depth track for NA12878 genome. Multiple SV pairs in the same locus were grouped together, thereby reducing the 31 candidate pairs to 16 candidate regions. Next, we extracted sequence reads which were longer than 2 kbp, soft-clipped more than 100 bp, and were mapped within a 1 kbp window, both upstream and downstream, of each boundary of a candidate region. We then used AGE[33] to realign the extracted reads around those loci. By inspecting AGE and BLAT alignments, we derived the exact SV breakpoints.

In this way, we determined the precise breakpoint boundaries for 9 candidate regions. Of these, three SVs were single deletions, three SVs represented two or three adjacent (independent) deletions, and three represented SVs with complex breakpoints. The complex cases are: (i) deletion with an insertion into the region outside of the boundary (chr4:91931666-919357970); (ii) double deletion, with an inversion (chr17:5594699-5597504); and a multi-deletion-inversion-duplication event (chr11:55365292-55457586, shown in Figure 3). For the remaining 7 candidate regions, the breakpoints could not be resolved. Of these, two regions resided in VDJ loci, where multiple split-reads suggested the presence of various breakpoints. In two additional cases, the distance between SVs in a pair was larger than 15 kbp, *i.e.*, beyond the span of PacBio or Moleculo reads. And in three cases we could not find any supporting split-reads. These SVs were in highly repetitive regions, where Moleculo and PacBio based analyses showed limitation (in the case of Moleculo due to the short underlying DNA sub-reads, and in the case of PacBio due to inherent high sequencing error rate).

Characterization of microhomology and mapping of template sites (contributed by: Alexej Abyzov, Taejeong Bae, Hugo Lam, and Jasmine Mu):

We compared the formation mechanisms of assembled breakpoint-resolved deletion SVs in this study with breakpoints from Conrad *et al*.[35] and Mills *et al.*[36], by examining sequence patterns of breakpoints based on BreakSeq[37] and manual review (**ED Figure 9**). The Mills *et al.* study based on 1000GP pilot data exhibited a comparably higher fraction of SVs with a repeat-associated formation mechanism (i.e. non-allelic homologous recombination, and mobile elements), which may be explained by the fact that in Mills *et al.* the 1000GP SV

analysis group focused on SV discovery only (rather than genotyping) and did not enforce a global FDR cutoff. In our current study, due to an emphasis on identifying SVs for which high confidence genotypes can be obtained, we do expect some bias against repetitive elements, with the fewer number of reads mapping to them interfering with genotype quality. Furthermore, the relatively small size of Alu elements led to relatively to fewer reads being mapped to them, resulting in less robust discovery and genotypes for this class too.

Mapping of template sites and characterization of microhomology (contributed by: Alexej Abyzov, Taejeong Bae): For each deletion site (≥50 bp) with an insertion (≥20 bp), we searched for the origin of the insertion site (*i.e.* insertion template[38]) using BLAT (http://www.kentinformatics.com). BLAT imposes a minimum 20 bps limit on the sequence length it attempts to align, consequently we conducted mapping for insertions of at least 20 bps in length. Each insertion was aligned to the hg19 reference genome using a BLAT web interface (http://genome.ucsc.edu/cgi-bin/hgBlat). We manually examined alignments with the aim of identifying a single template site for each insertion based on the alignment, such that (i) the insertions is aligned almost full length with few mismatches and/or short indels; (ii) the alignment has a considerably better alignment score than other alignments. Out of the 4,813 complex deletions described in the main text, 796 showed an insertion of at least 20 bp, and hence were used in this analysis. Out of these, we were able to unambiguously map exactly one template site for 441 complex SVs (ED Table 3-C). We determined micro-homologies (MH) between individual breakpoints of the deletion and the corresponding template site boundary (Supplementary 4.3.1). A random distribution was obtained by assessing MH between: (i) the left (denoted "5'") deletion breakpoint and right (denoted "3'") template site boundary; (ii) as well as between the right (denoted "3'") deletion breakpoint and left (denoted "5'") template site boundaries. Breakpoints partaking in complex SV sites exhibit a similar distribution of microhomology lengths as the bulk of non-complex (simple) SVs in our callset (**ED Figure 9)**. This pattern of microhomology observed between individual deletion breakpoints and corresponding insertion template site boundaries is consistent with formation of complex SVs by a single mutational event, presumably through template switching.

Analysis of SVs formed involving long homology stretches (e.g. by non-allelic homologous recombination; NAHR) (contributed by: Alexej Abyzov, Taejeong Bae, Hugo Lam, and Jasmine Mu):

To investigate SVs that based on BreakSeq were inferred to be formed by NAHR in more detail, the RepeatMasker track was downloaded from the UCSC Browser, comprising 5,298,130 repeat annotation entries for HG19, out of which 1,142,278 and 916,234 were Alu and L1 annotations respectively. 1,019,022 of the Alus and 22,124 of the L1s, which were larger than or equal to 150 bp and 3000 bp (>=~50% of their full lengths) respectively, were extracted to facilitate annotation of both the left and right breakpoints of deletion events. Deletions classified as NAHR for which both breakpoints annotated with the same type of repeat (*i.e.,* Alu or L1) were inferred to be repeat-mediated (*i.e.*, Alu-

mediated or L1-mediated). Out of 2,936 NAHR events inferred in our study, 1,777 (61%) are Alu-Alu-mediated whereas only 8 events (0.3%) were found to be L1-L1-mediated (the latter number is presumably a considerable underestimate due to ascertainment bias, given the inability of Illumina sequencing with short insert sizes in resolving this form of variation sensitivity). The remaining NAHR events were inferred to be formed by other repeat classes.

5. Validation Experiments

5.1. PCR validation of SV callsets

Experimental conditions of PCR validations: PCR experiments were carried out in different laboratories, focusing on different SV types: EMBL (DEL, DUP, INV, NUMT), LSU (MEI), and UMICH (NUMT).

EMBL

Contributed by: Adrian Stütz, Benjamin Raeder, Thomas Zichner, Tobias Rausch, Jan Korbel

PCR primers were obtained from Sigma. PCRs were performed using 10ng of genomic DNA (Coriell) in 25 μl volumes using the Sequalprep Long PCR reagents (Life technologies) in a 96 well plate using the DNA Engine Tetrade 2 thermocycler (BioRad). PCR conditions were: 94°C for 3 minutes, followed by 10 cycles of 94°C for 10 s, 62°C for 30s and 68°C for 6 minutes and 25 cycles of 94°C for 10s, 60°C for 30s and 68°C for 8min, followed by a final cycle of 72°C for 1 minutes. PCR products were analyzed on a 0.8% agarose gel stained with Sybr Safe Dye (Life Technologies) and a 100 bp ladder and 1 kb ladder (NEB). If necessary, gel bands were cut with a scalpel, gel extracted with the Nucleospin Gel and PCR Cleanup kit (Macherey-Nagel) and send for capillary sequencing (GATC Biotech AG).

Allele frequency adjusted random site selection: For PCR validations of SV callsets, we focused on a subset (rather than the entire set) of phase 3 samples. Validation sites were picked in this subset of samples in an allele-frequency weighted manner in order to avoid biasing PCR validations to common SV sites. First, an allele-frequency histogram was computed on all genotyped sites. Second, the site list was subsetted to SVs were at least one of the validation samples is a carrier and each site was annotated with its original allele frequency. Last, validation sites were randomly picked in each frequency bin using the annotated allele frequency of the validation sites and the proportion of sites in the original allele frequency histogram.

Primer design: To design PCR primer pairs for the validation of a given SV, we implemented a computational SV validation primer design pipeline in Python 2.7, made available at https://github.com/zichner/primerDesign (2015). The pipeline utilizes Primer3[39] as well as BLAST[40] and is based on the following steps. First, the pipeline extracts the genomic sequence of a 200 bp region next to each SV break point; the side and orientation of the regions are depending on the SV type (see below). Then, Primer3 is applied to compute a set of primer pairs for these regions. Subsequently, all primer sequences are tested for their uniqueness across the genome using BLAST. Primers are considered as unique only if all their off-target hits have at least four mismatches, or at least three mismatches if all of them are at the 3'-end of the primer. If at least one primer pair is found where one or both primers fulfill the uniqueness condition, the pair with the best Primer3 score is reported. Otherwise, the size of the regions for which primers are

designed is increased from 200 bp to 600 bp and the BLAST step is repeated. If this does not result in a valid primer pair, the analysis is repeated for a region of 2000 bp and afterwards 6000 bp. If still no primer pair can be designed, the corresponding SV cannot be considered for validation through this procedure.

***The following SV classes were systematically targeted for validation at** EMBL*:

***Deletions**:* a pair of primers was placed outside of/flanking the predicted SV. This will result in either a band with the expected size based on the reference genome, and/or a band smaller, corresponding to the deletion allele. The band pattern therefore allows distinguishing 0/0, 0/1 and 1/1 genotypes. In case of small deletions where no size difference will be observed, capillary sequencing will be used to confirm the presence of the small deletion.

***Tandem Duplications**:* a pair of primer is placed within the predicted SV in an outward facing orientation. This will result only in a band if a tandem duplication occurred, and therefore this procedure can verify the presence of tandem duplications, but cannot genotype them.

***Inversions**:* 4 primers are designed, primer 1 and 4 are flanking the SV, and primers 2 and 3 are inside of the SV. The reference allele will be seen as primer combinations 1+2 and 3+4, whereas the inversion allele will be seen as 1+3 and 2+4. All 4 tests need to be performed to be able to distinguish 0/0, 0/1 and 1/1 genotypes. It is of note that this procedure leads to considerable failure rates in the presence of highly complex inversions (such as those shown in **Figure 3** in the main text and **ED Figure 10**), which is why long-read-based targeted sequencing – which we found to be considerably more robust to inversion complexity – was performed as an additional validation step (see further below).

***NUMTs**:* a pair of primers was placed outside of/flanking the predicted SV. This will result in either a band with the expected size based on the reference genome, and/or a band larger, corresponding to the NUMTs insertion allele. The band pattern therefore allows distinguishing 0/0, 0/1 and 1/1 genotypes.

LSU

Contributed by: Miriam Konkel, Jerilyn Walker, Mark Batzer

***MEI**:* Randomly selected, allele frequency adjusted, MEI candidate loci (64 Alu, 61 L1, and 65 SVA loci) were included in our PCR validation analysis. For primer design, 600 bp of flanking sequence were added up- and downstream of the breakpoint (insertion) coordinate. The sequence was retrieved from the human reference genome [hg19] using Galaxy[41,42]. Prior to primer design, all sequences identified as Alu elements in the flanking sequence were masked to Ns using RepeatMasker[43]. A safety margin of 50 nucleotides up- and downstream of the insertion coordinate was granted for each candidate locus to account for imprecise breakpoint calling.

Primer pairs were selected using BatchPrimer3 v2.0[44]. Each primer was subjected to a BLAT[21] analysis. Primers showing more than one match in the human genome were redesigned with Primer3[45]. Prior to primer redesign, the repeat content of the flanking sequence was determined using RepeatMasker. To determine if the flanking sequence matched to highly homologous loci, the flanking sequence was queried against the human reference genome [hg19] using BLAT. In cases with high sequence homology, the most homologous sequences were retrieved using the UCSC genome browser[21]. Following an alignment of the candidate locus with the other orthologous loci using the ClustalW feature in the BioEdit program[46] primer design was performed in regions with unique sequence to the candidate locus with mismatches in other highly homologous loci. The primers were queried against the human genome using BLAT and an *in-silico* PCR was performed to confirm the presence of only one PCR product and the amplicon size, which represents the empty allele (insertion absent).

A second primer pair was designed for loci originally identified as false positive if re-analysis of the MELT calls either clearly showed the presence of an MEI insertion and/or suggested a breakpoint outside of the original sequence coordinates. In these cases primer design was performed with Primer3 including all steps for primer redesign (see paragraph above).

For the analysis of the L1 and SVA candidate loci, previously designed internal primers were utilized[18,36]. In case of L1s, the primers resided in the 3'-terminus of the L1 consensus sequence. For the analysis of SVA primers in the 3' (up to five primers) and if necessary in 5'-terminus (up to three primers) were used. All PCR primers were ordered from Sigma Aldrich, Inc. (St. Louis, MO). The PCR primer sequences used in this validation study are available at http://batzerlab.lsu.edu (2015).

***DNA Samples for PCR verification*:** A subset of 24 DNA samples from Phase III and the YRI trio were used for the PCR validations (Table 5.1.1). The DNA panel also included human cell line DNA (HeLa; ATCC CCL-2) as well as "Pop80", a locally pooled DNA sample from different individuals of diverse geographic origins (Asia, Africa, South American, and European). As another PCR control, chimpanzee DNA (NS06006, Coriell) was included on the panel, representing the presumptive pre-insertion site (empty site) for each MEI event in human.

***PCR details**:* Using either a Perkin Elmer GeneAmp 9700 or a BioRad i-cycler thermo-cycler, PCR amplifications were performed in 25 μl reactions in a 96-well format. Each PCR reaction contained 15-25ng of template DNA; 200 nM of each oligonucleotide primer; 1.5 mM $MgCl_2$; 1X PCR buffer (50 mM KCl; 10 mM TrisHCl, pH 8.3); 0.2 mM dNTPs; and 1-2 U *Taq* DNA polymerase.

Candidate loci containing putative Alu elements were amplified using external primers (*i.e.*, primers flanking the mobile element insertion). In cases of no amplification of the candidate Alu element in the predicted individual, a temperature gradient PCR was performed to optimize the annealing temperature of the reaction and then a "hot-start" PCR was performed using Jumpstart *Taq* DNA polymerase (Sigma Aldrich, St. Louis, MO).

For L1 and SVA candidate loci, a minimum of two separate PCR reactions were performed to amplify the filled and the empty amplicon. To determine the absence of an insertion, a PCR was performed using the external primer pair. A second PCR used one primer residing within the L1 or SVA insertion (internal primer) in conjunction with an external primer (forward or reverse, depending on the orientation of the predicted insertion) to amplify the filled site. Moreover, all L1 candidate loci were subjected to the amplification of the entire L1 in at least one individual using the external primers and Takara LA-*Taq* (Clontech Laboratories, Inc., Mountain View, CA), a long-range polymerase. This was done in order to determine the size of the L1 insertion and/or to determine if the L1 was indeed not present if the putative L1 was not amplified in the predicted individual using the internal PCR approach. In the latter case, the long range PCR was performed on the whole DNA panel. (SVA candidate loci were not analyzed using a long range PCR approach because these PCR reactions are commonly unsuccessful due to the high GC-content, the length, and the highly variable number of tandem repeat (VNTR) region.)

PCR reactions were performed under the following conditions using a standard *Taq* polymerase: initial denaturation at 94°C for 90 s, followed by 32 cycles of denaturation at 94°C for 30 s, annealing at 57°C and extension at 72°C for 30 to 90 s depending on the predicted PCR amplicon size. PCRs were terminated with a final extension at 72°C for 2 min. For the amplification of the entire L1 using LA-*Taq* DNA polymerase, the above-described protocol was modified in the following way. The extension step of each cycle was carried out at 68° for 8 min 30 s, followed by a final extension step at 68° for 10 minutes at the end of the run. All PCR products (20 μl) were size-fractionated in a horizontal gel chamber on a 2% or 1% (for loci amplified with LA-taq) agarose gel containing 0.1 μg/ml ethidium bromide for 45-60 min at 175-200V or 1 hour/45 min at 150V, respectively. DNA fragments were visualized with UV-fluorescence and images were saved using a BioRad ChemiDoc XRS imaging system (Hercules, CA).

In addition to the estimation of the false detection rates, genotypes were recorded for each individual and locus. The genotypes were determined based on the predicted amplicon size of the empty site and the size-fractioned PCR products on the agarose gels.

Table 5.1.1:

Sample	Population	Population Description	Gender
HG00096	GBR	British from England and Scottland	Male
HG00268	FIN	Finnish from Finland	Female
HG00419	CHS	Han Chinese South, China	Female
HG00759	CDX	Chinese Dai in Xishuangbanna, China	Female
HG01051	PUR	Puerto Rican	Male
HG01112	CLM	Columbian in Medellin, Columbia	Male
HG01500	IBS	Iberian Populations of Spain	Male
HG01565	PEL	Peruvian in Lima, Peru	Male
HG01583	PJL	Punjabi in Lahore, Pakistan	Male
HG01595	KHV	Kinh in Hochi Minh City, Vietnam	Female
HG01879	ACB	African ancestry from Barbados in the Caribbean	Male
HG02568	GWD	Gambian in Western Division, The Gambia	Female

HG02922	ESN	Esan from Nigeria	Female
HG03006	BEB	Bengali in Bangladesh	Male
HG03052	MSL	Mende in Sierra Leone	Female
HG03642	STU	Sri Lankan Tamil in the UK	Female
HG03742	ITU	Indian Telugu in the the UK	Male
NA18525	CHB	Han Chinese in Beijing, China	Female
NA18939	JPT	Japanese in Tokyo, Japan	Female
NA19017	LWK	Luhya in Webuye, Kenya	Female
NA19625	ASW	African ancestry in Southwest USA	Female
NA19648	MXL	Mexican ancestry in Los Angeles, CA (USA)	Female
NA20502	TSI	Toscani in Italia (Tuscans in Italy)	Female
NA20845	GIH	Gujarati Indians in Houston, TX (USA)	Male
NA12878	CEU	CEPH Utah, USA	Female
NA19238	YRI	Yoruba in Ibadan, Nigeria; Mother of trio	Female
NA19239	YRI	Yoruba in Ibadan, Nigeria; Father of trio	Male
NA19240	YRI	Yoruba in Ibadan, Nigeria; Daughter of trio	Female

UMICH
Contributed by: Sarah Emery, Jeffrey Kidd

NUMTs*:* NUMTs identified by computational analysis were validated by polymerase chain reaction (PCR) and Sanger sequencing of amplicon(s) that spanned 50-500 bp of gDNA flanking the insert, the breakpoint between the gDNA and the insert, and the insert. Primer sets that hybridize to the gDNA flanking the insert were designed using Primer3 Software (http://www.genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi, [2015]) and amplification was done with Platinum Taq (Invitrogen Life Technologies, Gaithersburg, MD), Picomaxx (Agilent Technologies, Palo Alto, CA), or LongAmp (New England Biolabs, Beverly, MA) products in a 20-50ul reaction volume containing 50 ng of template DNA, 1 uM primer, and 1.5mM $MgCl_2$ if not supplied in the PCR buffer. Thermocycling was done for 30 cycles at 56-67 ºC annealing temperature and 1-15 minute extension time. For inserts less than 3 kbp, a PCR product of the predicted size was identified in individuals homozygous or heterozygous for the insert by agarose gel electrophoresis and the insert was sequenced in one individual. Amplicons of interest were purified from a PCR reaction for homozygous individuals (Qiaquick PCR purification kit, Qiagen, Valencia, CA) or isolated from the gel for heterozygous individuals (Qiaquick Gel Extraction Kit, Qiagen) and sequenced at the University of Michigan Sequencing Core. For inserts larger than 3 kbp, a PCR product of the predicted size was identified in individuals heterozygous for the insert by gel electrophoresis. For sequencing, two overlapping PCR products were made using primer sets designed as outlined above with one primer that binds in the gDNA flanking the insert and one primer that binds in the middle of the insert.

5.2. Long-read-based validation of inversions

Targeted validation of inversions using long DNA read data*:* Experimental inversion validations using targeted long-read (*i.e.*, PacBio or Oxford Nanopore MinION) sequencing of PCR amplicons were performed at two centers (EMBL Heidelberg and Baylor College of

Medicine). Targeted PacBio sequencing of fosmid clones was pursued at the University of Washington.

EMBL Heidelberg
Contributed by: Tobias Rausch, Markus Fritz, Adrian Stütz, Sascha Meiers, Andreas Untergasser, Jan Korbel

Predicted genomic inversions: Inversions inferred to be present in sample NA12878 were verified in two independent long read data sets: a) high coverage NA12878 PacBio data generated at Mount Sinai Hospital and b) high coverage NA12878 Illumina Moleculo data generated by Illumina.

PacBio and Moleculo reads were aligned to the hg19 reference using the BLASR[30] read aligner. For every predicted inversion in NA12878 all reads were extracted spanning the entire inversion locus. MUMmer-3 (ref[47]) was used to compute forward and reverse matches between the sequence read and the reference slice it aligns to. Matches were plotted and plots were manually screened for a "diagnostic inversion signature" (see e.g. example inversions shown in Figure 3).

Primer design + PCR: Primer design for 96 randomly selected, frequency adjusted inversions between 1-3 kbp in size was done as described above with the following modification: primers were placed at least 1 kbp away from the predicted breakpoints to allow unbiased amplification of both alleles even in the presence of an additional flanking deletions/insertions (*i.e.*, appreciable "complexity"). PCR amplicon sizes were between 3-9 kbp.

PCR primers were obtained from Sigma. PCRs were preformed using 10ng of genomic DNA (Coriell) in 35 μ l volumes using the Sequalprep Long PCR reagents (Life technologies) in a 96 well plate using the DNA Engine Tetrade 2 thermocycler (BioRad). PCR conditions were: 94° C for 3 minutes, followed by 10 cycles of 94° C for 10 s, 62° C for 30s and 68° C for 8 minutes and 25 cycles of 94° C for 10s, 60° C for 30s and 68° C for 10 minutes, followed by a final cycle of 72° C for 10 minutes. 10ul PCR product aliquots were analyzed on a 0.8% agarose gel stained with Sybr Safe Dye (Life Technologies) and a 1 kbp ladder (NEB).

The band pattern, size and intensity was recorded and grouped into three classes: I. PCR failures/unspecific reactions (11X); II. strong (36x) and middle (32X) intensity bands and III. weak (10X) and very weak (7X) intensity bands. Next, 5ul of the remaining PCR of strong band loci and 8.75ul of middle bands (1:1.75 ratio) were mixed and purified by adding 237.5ul of AMPure XP beads (Agencourt, 0.5X volume) and eluted in 50ul nuclease free water. The same was done for 4ul of weak bands and 20ul of very weak bands (1:5 ratio) followed by AMPure XP bead cleanup (0.5X). The concentration of amplicon PCR pools was quantified with the Qubit BR kit (Life Technologies).

Pacific Biosciences Amplicon library prep + sequencing*:* The purified amplicon PCR DNA pool was used with the 2 kbp template preparation and sequencing protocol (Pacific

Biosciences) with slight modifications such as inclusion of the repair DNA damage step from the 6 kbp template protocol. Briefly, 2.5ug of amplicon pool DNA was cleaned up with provided special AMPure PB beads (0.6 volumes) and eluted with a concentration of 60 ng/µl. Afterwards, the repair DNA damage step was performed on ice by individually adding 5 ul DNA damage repair buffer, 0.5 ul NAD+, 5 ul ATPhigh, 0.5 ul dNTP and 2 ul of DNA damage repair mix in 50 ul and incubating it for 20 minutes at 37°C. After addition of 2.5ul End repair mix and 5 minutes incubation at 25 °C, an AMPure PB cleanup step (0.6 volumes) was performed and eluted into 30 ul provided elution buffer. Blunt adapter ligation reagents were individually added and incubated for 15 minutes at 25°C and heat inactivated for 10 minutes at 65°C. After addition of 0.5 ul ExoIII and ExoVII enzyme, the mix was incubated for 1 hour at 37°C. After two rounds of AMPure PB cleanup (0.6 volumes), the final library was eluted into 10 ul elution buffer and quantified with both Qubit HS (Life Technologies) and Bioanalyzer 12000 (Agilent Technologies). Further processing of the library and sequencing was done as recommended by the manufacturer and each library was sequenced on one SMRT cell using P4-C2 chemistry at the core facilities of the Max-Planck Institute (Köln, Germany).

Oxford Nanopore MinION library prep + sequencing*:* The purified amplicon PCR DNA pool was used with the genomic DNA sequencing kit (version SQK-MAP002) for MinION library prep as part of the MinION early access programme (Oxford Nanopore Technologies). Briefly, 1.5-2 ug of amplicon pool DNA and 5 ul of DNA-CS were end repaired using the end repair module reagents (NEB) for 30 minutes at 20°C, purified with 0.5 volumes of AMPure XP beads and eluted in 25.2 ul nuclease free water. A-tailing (NEB) was performed in 30µl for 30min at 37°C and followed by adapter ligation (Oxford Nanopore Technologies) by adding 10 µl adapter mix, 10 ul of HP adapter and 50 ul of blunt T/A ligase mix (NEB) and incubation for 10 minutes at 20°C. A special AMPure XP cleanup step (0.4x volume) was performed, using 150 µl of provided wash buffer instead of 70% EtOH once and elution into 25 µl provided elution buffer without a drying step. Next, tether annealing was performed by adding 10 µl tether mix and incubation for 10 minutes at 20°C and followed by the library conditioning step by addition of 15 ul HP motor mix and incubation o.n. at 20°C at 750 rpm. Briefly before the MinION sequencing run, 6 µl of prepared library was mixed with 140 µl EP buffer and 4 µl of fuel mix, gently mixed to produce the final library and loaded on a primed MinION flowcell (version FLO-MAP001 and FLO-MAP002). MinION flowcells were analyzed with the software client Metrichor v 0.17.39962, the sequencing software MinKNOW v 0.46.1.9 and the 2D workflow v1.7. We considered flowcells with more than 200 active pores in the MAP_Platform_QC run. Flowcells were primed with 150 µl EP buffer followed by 10 minutes waiting time, before 150 µl of final amplicon library were loaded and sequencing was initiated.

Comparison of PacBio versus MinION inversion validation data: A subset of 69 PCR amplicons was analyzed using both PacBio as well as MinION technology, with the primary goal to assess the potential of Oxford Nanopore MinION sequencing to verify and classify inversions. In 37 cases (53.6%), both technologies agreed in revealing the presence (35 loci) or absence (two loci) of inversions in given samples, and additionally in each case the inversion was inferred as present both techniques agreed in the classification of the inversion type. Example plots for each inversion class are shown below. In the remaining

cases, 20 loci could be reliably analyzed by PacBio data but remained uninformative in MinION data due to noisy data or low coverage, whereas two loci could be analysed with MinION data but not with PacBio data. Thus, in spite of overall higher read error rates seen for the MinION technology, both long-read technologies were deemed to be suitable for verification and characterization, with classifications showing 100% concordance for loci with sufficient coverage.

The FDR estimate for inversions based on amplicon PCR sequencing is likely conservative for the following reason: We could not exclude allelic dropouts of the inversion allele for particularly complex inversion loci where PCR primers were unable to anneal, and accordingly observed a strong allelic bias of the variant allele relative to the reference allele for several loci (up to 1:100 for variant allele vs. reference allele), presumably since the inversion allele was disadvantaged during the PCR step or during primer annealing.

Baylor College of Medicine
Contributed by: Fuli Yu

To enable validation of inversions and NUMTs with PacBio reads, unique variant sites were selected for experimental validation using the BCM-HGSC long-range PCR amplification and PacBio sequencing pipeline, using amplicons 3-4 kbp in size. After PacBio library preparation, three PacBio libraries were individually sequenced per SMRT cell following the manufacturer's Guide – Pacific Biosciences Template Preparation and Sequencing, version 10.

Both the circular consensus sequences (CCS) reads and continuous long reads (CLR) were mapped against human reference genome GRCh37. We used BLASR[30] and BWA-SW[48] aligners to verify performance, and selected the CCS BLASR pipeline for data processing. For validation, we manually inspected each amplicon per sample site using the IGV browser (results were generally concordant between CLR and CCS reads).

University of Washington
Contributed by: Maika Malig, Mark Chaisson, Evan Eichler

We selected a total of 35 inversion sites (inferred by DELLY) from two genomes (NA12756 and NA19129) for validation using long-read (PacBio) SMRT sequencing of fosmid clone inserts (~40 kbp). A total of 113 clones (2-4 clones per site) were selected and grown based on mapping of fosmid end-sequence pairs to GRCh37[49]. DNA was individually prepared for each clone (High Pure Plasmid Isolation Kit™, Roche) and DNA from 7-8 clones were pooled. A 20 kbp SMRTbell™ template library was prepared for each pool; the library sequenced with one SMRTcell per pool using either P4-C2 or P5-C3 chemistry and inserts were assembled using HGAP and QUIVER post-processing[31] as previously described[50]. 111/113 (98.2%) of the clone inserts resolved into a single se quence contig with on average 400-fold sequence coverage per fosmid clone insert. Assemblies were compared with GRCh37 using Miropeats[51] and dotplot analysis to identify breakpoints and confirm inversion status. Overall, 82.3% (28/34) of sites validated with 1 site excluded due

to sequence complexity. This is a conservative estimate because only one haplotype was recovered for 2/6 of the invalidated sites. Excluding these two sites would result in a validation rate of 87.5%. We further employed PacBio reads from the recently sequenced CHM1 genome[50] for the verification of phase3 inversions which the EMBL group genotyped into CHM1 using published CHM1 Illumina sequencing data[52].

5.3. Validation of CNVs using IRS

Contributed by: Bob Handsaker, Seva Kashin, Peter Chines, Tobias Rausch

The IRS test estimates a FDR for a set of putative copy number unbalanced (or CNV) SV calls – *i.e.*, deletions, duplications, and mCNVs – by utilizing the distribution of a test statistic (across all calls) derived from the relative probe-level intensities of the same probe(s) between samples expected to have different copy number levels. We utilized SNP probe intensities from two different SNP arrays: the Omni 2.5 and the Affymetrix 6.0. The probe intensities were normalized and summarized as described below.

Using the normalized probe intensities, for each genotyped SV site, up to two tests were performed: One test based on samples with predicted copy number less than the reference copy number of two copies (dels) and one test for samples with predicted copy number greater than two (dups). For the first test, the samples are divided into two subsets, those with predicted copy number two and those with predicted copy number less than two (other samples are not used in the first test). For each probe underneath the SV, the samples are first ranked according to the probe intensities with ties broken randomly. Then using the ranks at each probe, the samples are re-ranked across all probes, with ties broken randomly. A rank-sum test is performed to test whether the samples predicted to have copy number below the reference copy number have lower ranks than the samples with reference copy number. The second test is symmetrical to the first test, comparing the subset of samples with copy number above two to the samples with reference copy number.

An implementation of this test is available as the IntensityRankSum annotator module in the Genome STRiP software.

For each SV, these tests yielded either one or two p-values depending on the range of copy number genotypes at that SV. The FDR of a set of SVs was estimated by dividing the putative SVs into three subsets: (a) SVs with observed copy numbers either at or below the reference copy number (b) SVs with observed copy numbers either at or above the reference copy number and (c) SVs with observed copy numbers both above and below the reference copy number. Subsets (a) and (b) have one p-value while subset (c) has two p-values. For subsets (a) and (b), we estimate the FDR of these subset as two times the fraction of sites with p-value >0.5. For subset (c) with two p-values, we estimate the FDR of this subset as four times the fraction of sites having both p-values >0.5. An overall FDR for the call set as a whole is calculated as the weighted sum of the FDRs of the three subsets (a-c). The final IRS FDR estimates for the bi-allelic deletions, bi-allelic duplications and multi-allelic copy number variants (mCNVs) are shown in Table 1 (main text).

Generation of Array Intensity Matrices for IRS test: The probe intensity values for the Affymetrix 6.0 array were generated from array data generated and contributed by Coriell. Array data was available for 2,504 samples, of which 2,476 were included in the final 1000 Genomes data set.

For the Affy 6.0 array, the probe intensities were first normalized using the apt-probeset-summarize utility from Affymetrix with the following parameters:

apt-probeset-summarize
--cel-files ***cel-file-list***
-a "quant-norm.target=1000,pm-only,plier.optmethod=1,expr.genotype=true"
--cdf-file ***cdf-file***
--probeset-ids ***probeset-id-list***
--precision 2

The normalized intensity values from the A and B SNP probes were then summed (for the Affy 6.0 copy number probes, the individual probe intensity was used).

The probe intensity values for the Illumina Omni 2.5 array were generated from arrays run at the Broad Institute for 2,141 samples, of which 1,639 were included in the final 1000GP data set.

For the Omni 2.5 array, normalization of the probe intensities was performed at the Broad Institute using the default protocol for SV analysis for the BirdSuite software package[53] using the InfiniumIDATParser utility. The probe intensities for the A and B SNP probes were then summed.

For the Omni 2.5 array, all of the array probe sequences were realigned using the protocol specified below and probes with a score less than 13 were excluded. The realigned coordinates were used to determine the start and end position of each probe against the reference genome.

BWA (version 0.5.8c) was used to align the probe sequences to GRCh build 37 of the human genome, with default parameters. Most Infinium assays use a single 50 bp probe, and assay the position immediately 3' of the annealed probe by a single-base extension reaction. Assays of A/T and C/G SNP assays use two different probes that differ at only the last position, which is the nucleotide to be assayed. For the latter assays, we first removed the last base of the probe, and then mapped the remainder of the probe in the same manner as the other assays. Thus the probe regions for the two-probe assays are 49 bp long.

The BWA alignments are processed to identify the exact position being assayed, the orientation of the assayed alleles with respect to the reference, the distance of the nearest mismatch, if any, from the 3’ end of the probe, and the number of next-best hits with a single additional difference from the reference genome. If any variant from the 1000 GP Phase 1 integrated panel (minor allele frequency >1%) is found in the region where the probe aligned, the variant nearest to the 3’ end of the probe was noted.

In addition to annotating each assay with the above characteristics, we also assigned a score, based on a ranking of the "most deleterious" annotation for each assay:

Score	N assays	Description
0	440	unmapped (not included in file)
1	269	probe aligns to more than two loci (the BED file shows one of these, chosen at random)
2	1,579	probe has more than three next best hits, or next best hit does not differ within 7 bp of 3' end
10	157	mismatch within 7 bp of 3' end of probe
11	2,819	probe aligns to two loci
12	2,530	probe has three or fewer next best hits, all of which have a difference within 7 bp of the 3' end
13	37,550	variant within 7 bp of 3' end of probe
30	515	indel more than 7 bp from 3' end of probe
35	4,443	mismatch more than 7 bp from 3' end of probe
38	27	neither allele matches reference
40	296,741	variant more than 7 bp from 3' end of probe
44/46	948	additional hit to "random" contig, not a chromosome
50	2,101,982	no issues detected

5.4. Validation of CNVs using array-CGH

Contributed by: Ankit Malhotra, Tobias Rausch, Chengsheng Zhang, Dariusz Plewczynski, Kamen Radew, Eliza Cerveira, Mallory Romanovitch, Przemyslaw Szalaj, Ryan Mills, Charles Lee

We designed a custom Agilent 1M CGH microarray (aCGH) for validation purposes, targeting known structural variation sites from various sources including the 1000GP Pilot[36] and Phase 1[54] releases, DGV[55], and other recent publications[9,56] for a total of 22,531 deletions, 46,268 duplication; 4,873 MEIs and 142 retroduplications. These variants were segmented into discrete regions of overlap between overlapping events, with between 1 to 7 custom probes assigned to each individual segment. Remaining probes were uniformly distributed across the genome to create a “backbone”. We obtained genomic DNAs of the phase 3 samples from the Coriell Institute for Medical Research and performed CGH using the standard protocol provided by the manufacturer. Briefly, the testing DNA sample and the reference DNA sample (NA10851) were fragmented by enzymatic digestion with Alu1 and Rsa1, and labeled with Cy5 and Cy3, respectively, followed by co-hybridization to the custom microarrays. After hybridization, arrays were washed and scanned, and the final feature extraction files were used for data analysis.

We generated aCGH data for all individuals from the phase 3 set. To control for noise, we excluded probes that were inconsistent across the whole population of individuals as follows. For each probe, we first calculated the mean and standard deviation of the

reference channel intensity across the entire set of arrays. As the reference signal is coming from the same sample (NA10851), we expect each individual probe to behave similarly across essentially 2,500+ replicate experiments. All probes whose reference intensity value fell greater than 1 standard deviation away from the mean of all probes were excluded. A similar filtering was applied using the log2 ratio of the sample/reference signal. However, here we allowed a lower bound of 2 standard deviations before excluding probes. We also limited our analysis to internal probes only, defined as probes falling completely within the bounds of the interrogated region.

Next, we corrected individual arrays for %GC bias and normalized them across all the individuals to their respective medians. To correct for the %GC bias, we binned probes from each array into 14 %GC content bins (from less than 20% GC content to greater than 80% GC content). Each bin was then centered to a mean log2 ratio of 0. To normalize arrays from for population level analysis, the median log2 ratio of each array was then centered to 0.

We next developed a custom software named *canny* (https://bitbucket.org/remills/canny [2015]) to assign integer copy numbers (CN) to each predicted variant for each sample. Briefly, this software first intersects the collection of filtered probes derived above with individual deletions and duplications from each algorithm in order to assign sets of probes to each individual predicted variant. Regions with >5 probes are retained and the median value is calculated across each region for every sample. These are then clustered into discrete copy number states using a mean-shift approach (R package: LPCM) with bandwidth=0.05 and threshold=0.3. The largest cluster was set to CN=2 with neighboring clusters sequentially lower or higher, respectively, and were removed entirely if the read-depth of the reference sample (NA10851) was significantly higher or lower than its GC-corrected mean coverage indicating a skewed baseline copy number ratio.

The aCGH derived copy number for each predicted variant and each sample was used to guide the merging of structural variant predictions from individual tools. In particular, the aCGH data confirmed the assumption that paired-end genotyping methods are inclined to underestimate the deletion carrier samples in low-coverage sequencing data. RD based methods such as GenomeSTRiP, however, showed overall a high concordance between the sequencing derived copy number and the aCGH copy number in a set of manually inspected deletions. As a result, we decided to give priority to GenomeSTRiP genotyped deletion calls in a given set of redundant deletion calls. RD methods, however, can be difficult to calibrate to the true baseline copy number 2 state and hence, we also observed overlapping CNVs were the genotypes between independent prediction methods appeared to be copy number shifted. We again manually inspected such cases using the aCGH data to get an independent assessment of these copy number shifted sites. This analysis clearly suggested a preferential selection of those predictions where most samples were genotyped as copy number 2.

5.5. Validation of CNVs using Complete Genomics

Contributed by: Goo Jun

We developed a pipeline to genotype copy number unbalanced SVs (*i.e.*, CNVs) based on Complete Genomics (CG) WGS data, to enable comparing CG-based SVs to our phase3 dataset. This pipeline consists of two main steps: 1) identification of candidate intervals and merging overlapping intervals, as well as 2) multi-sample clustering and genotyping.

The first step is identification of candidate intervals. CG variant data include two different sources for possible deletion events: SV events from depth-based ploidy calls along 2,000 bp intervals, and so-called junction events detected from read mapping (a junction event denotes the case where a read is mapped across two separate regions of the reference genome; e.g. owing to a large deletion). We first collected junction intervals and SV intervals to build a single list of candidate intervals across all samples, then all candidate intervals were sorted first by starting positions and then by ending positions. All duplicate intervals were removed to avoid unnecessary computation. We used hierarchical agglomerative clustering by using R) as a similarity measure to obtain merged intervals. Within the set of overlapping intervals, two intervals with maximum RO were merged to a single interval, and then the process was repeated until there remained no pair of intervals with RO >0.5. Once the list of candidate intervals was finalized with no more significantly overlapping intervals, we collected the normalized average depth (GC-corrected) from each sample for each candidate interval.

The next step was to decide whether each interval is polymorphic (or not) by estimating Gaussian mixture models with 1, 2, and 3 components. The 1-component model is a single Gaussian distribution; hence the maximum-likelihood parameters are easily calculated by taking the mean and the variance of the data. Each component in the 2 or 3 component models has three parameters: (α, μ, σ), where α is the mixture weight, μ is the mean and σ is the standard deviation. These parameters are obtained by standard expectation-maximization algorithm. Each component in the mixture model is initiated with suitable values for 0, 1, and 2 deletions. Decision on number of components is based on Bayesian information criterion (BIC). BIC is a function of log-likelihood, number of components, and number of samples. B_k, BIC for k-component model is calculated as

$$B_k = -2\,\mathrm{LLK} + 2k \log n$$

, where n is the number of components in the model, k is number of free parameters in Gaussian mixture model and LLK is the overall log-likelihood of data given the model. We choose the model with the lowest BIC value. If either 2 or 3 component model is chosen, the next step is evaluating how well the components are separated.

Within Bayes decision rule, it is known that Bhattacharyya coefficient (BC) sets an upper bound on the theoretical classification error; hence we use BC as a measure for overlap between Gaussian components. The Bhattacharyya distance D between two Gaussian distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ is given as:

$$D = (\mu_1 - \mu_2)/8\ \sigma_{avg}^2 + 0.5 \log (\sigma_{avg}^2/\sigma_1\sigma_2)$$

, where $\sigma_{avg}^2 = (\sigma_1^2 + \sigma_2^2)/2$. BC is obtained easily from D as BC = $\exp(-D)$. If the model has more than 1 component and the maximum overlap between components (BC) is lower than the given threshold, then we conclude that the candidate interval contains deletions. The final step is making calls based on the posterior probability.

We genotyped 433 Complete Genomics sequenced samples. This dataset included 129 trios and 12 duos from 427 samples, and 6 samples that were sequenced twice with CG technology. Each sample was sequenced at >40× coverage. Sequencing data and variant information have been generated by Complete Genomics pipeline version 2.2. We identified 8,321 large deletions with maximum BC threshold of 0.01. Genotypes were set to be missing when posterior probability from the mixture model is less than 0.9, and average call rate is 99.9%. Because both trios and duos were represented in the data, it was possible to check for Mendelian consistency to measure genotyping accuracy. We used Merlin[57] to estimate genotype error rates. CG SV set showed extremely low error rate estimates of 0.1%.

6. Analysis of Structural Variation

6.1. Population genetic analyses

Deletions, Duplications, mCNVs
Contributed by: Peter Sudmant, Evan Eichler

Population Diversity*:* To assess the relative diversity of each of the individual populations assessed in this study, we calculated the per-individual SV-heterozygosity and SV-homozygosity for deletions, bi-allelic duplications, and multi-allelic copy mCNVs. We define SV-heterozygosity to be the total number of heterozygous events identified in an individual and SV-homozygosity to be the total number of homozygous events (**ED Figure 5**).

We find that African populations exhibit on average 27% more heterozygous deletions per individual than other populations (mean of 1,705 vs. 1,342), commensurate with the increased diversity of individuals from the African continent. We note that these differences are likely due to a shift in the allele frequency spectrum as has been noted for SNPs [58,59]. Next to Africans, Puerto-Ricans exhibited the highest deletion SV-heterozygosity, consistent with African-admixture into Puerto-Rican populations. Africans from the South West exhibited the largest variance among individuals, consistent with this population being made up of a diverse collection of admixed individuals. East Asian populations exhibited the lowest levels of deletion SV-heterozygosity. In contrast to deletion SV-heterozygosity, African populations exhibited the lowest levels of deletion homozygosity, again, consistent with the increased diversity of African populations. East Asian populations exhibited the highest levels of homozygosity.

PCA*:* To explore relationships of the different populations explored in this study we performed principal component analysis (PCA) using deletion genotypes. Genotypes were normalized as described by Patterson[60]. Briefly, for a particular deletion with 0,1,2 genotypes, the vector of genotypes was first centered about the mean, μ ,and then divided by$\sqrt{p(j)(1-p(j)}$where p(j) is an estimate of the allele frequency of a particular site, estimated as $\mu/2$. The resulting matrix of all transformed genotypes was then used for estimation of the principal components. PCA was performed on the entire set of populations, and additionally on each continental group independently. While PC1-4 described population structure when all individuals were included in the analysis, only PCs 1 and 2 exhibited structure in each of the continent-specific analyses, with the exception of the analysis of African populations.

PCA of all individuals separated African from non-African populations along PC1, and highlighted admixed populations. PC2 separated the European, South Asian and East Asian populations into distinct clusters while populations of the Americas clustered either intermediate to Europeans and South Asians (PUR, CLM) or intermediate to East Asians and South Asians (PEL, MXL). PC3 cleanly separated South Asians from all other individuals and PC4 described a long cline of all the Americas individuals with PUR and CLM clustering

closest to African, European and East Asian individuals and Peruvians and Mexicans (PEL, MXL) stretching the furthest from the African, European, East Asian cluster.

The African continental population-specific PCAs describe the proportion of admixture of ASW and ACB individuals along PC1 and separate into three clusters along PC2 made up of the Gambian (GWD) and Mende in Sierra Leone (MSL) individuals, the Yoruban (YRI) and Esan (ESN) Nigerian populations, and the Luhyan individual (LWK). The admixed ASW and ACB populations cluster closest to the Nigerian populations. PC3 further distinguishes LWK from the other populations with some overlap with GWD individuals.

The American continental population-specific PCA shows a cline of separation of Colombian (CLM), Mexican (MXL) and Peruvian (PEL) populations along PC1 with Puerto Rican populations separating along PC2.

The East Asian continental population-specific PCA separates Dai (CDX), Kinh (KHV), Southern Han (CHS), Han in Beijing (CHB), and Japanese (JPT) populations out long PC1. Little clustering is observed along PC2.

The European continental population-specific PCA largely separates Finnish populations from the remaining individuals with Toscani (TSI) and Iberian (IBS) individuals clustering together opposite the Finns. Again little to no clusters are observed along PC2.

The South Asian continental population-specific PCA largely exhibits two clusters along PC1 - a tight clustering of Sri Lankan (STU), Telugu from the UK (ITU) and Punjabi from Pakistan (PJL) populations and a looser cluster of Gujerati from Texas (GIH) and Bengali from Bangladesh (BEB).

These results are broadly similar to previous reports by Jakobsson *et al.* [61] and others.

***Vst analysis**:* To systematically identify SVs that show stratification among populations, we calculated Vst[62] for each structural variant. Vst is a metric that compares the variance between two populations of size n_1 and n_2 individuals respectively:

$$\frac{\sigma_T^2 - \frac{n_1 \cdot \sigma_1^2 + n_2 \cdot \sigma_2^2}{n_1 + n_2}}{\sigma_T^2}$$

where $\sigma^2$is the variance in copy number genotypes. For mCNVs, duplications and deletions, we thus calculated the Vst among all pairwise population and super-populations classifications. We selected a Vst cutoff of 0.2 to indicate population stratification of a locus, which has previously been used as a metric of high-stratification[9], identifying 16 highly stratified duplications, 2925 highly stratified deletions and 231 highly stratified mCNVs (**Table 6.1.5**).

Table 6.1.5. Total number of high-Vst events identified for each class and the number of gene intersecting high-Vst events using Vsts calculated among all sub-populations. In parentheses are the number of events identified for super-populations. X-linked calls are excluded.

SV type	total events	Vst>=0.2	Vst>=0.2 and intersecting genes
DUP	6120	16 (12)	10 (9)
DEL	42441	2872 (1312)	1113 (512)
mCNV	2994	227 (111)	113 (60)

MEIs

Contributed by: Eugene J. Gardner, Scott E. Devine

MEI population genetics studies

The MEI population genetics studies depicted in **ED Figure 5** were performed with the integrated phase 3 MEI call set. Missing genotypes were imputed using the phase 3 SNP/indel/large deletion haplotype scaffolds. The subset of 13,678 MEIs that were in HWE (p = 0.05)) was used in all MEI population genetics studies (ED Table 4). MEIs were required to be in HWE in all 26 populations to be included in the HWE set used for population genetics studies. This set included 10,378 *Alu*, 2,603 L1, and 697 SVA sites (52.0M, 13.0M, and 3.5M genotypes, respectively). Only MEIs on autosomes were used.

Phylogenetic Tree Construction

The phylogenetic tree depicted in **ED Figure 5** was constructed from twenty populations belonging to the four non-admixed super-populations EUR (CEU, FIN, GBR, IBS, TSI), SAS (GIH, PJL, BEB, STU, ITU), EAS (CHB, JPT, CHS, CSX, KHV), and AFR (YRI, LWK, GWD, MSL, ESN) using PHYLIP version 3.96[63]. Six populations from the AMR (CLM, MXL, PEL, PUR) and AFR (ACB, ASW) continental groups were excluded due to high levels of admixture in these populations.

The phylogenetic tree depicted in **ED Figure 5** was constructed as follows: First, MEI allele frequencies were determined in each population. Next, the GENDIST algorithm provided with PHYLIP was applied using the Cavalli-Sforza[64] genetic distance measurement. Finally, NEIGHBOR[65] was run with the GENDIST output in UPGMA mode to generate the neighbor joining tree depicted in **ED Figure 5**. To generate bootstraps, 100 replicates were performed as outlined above with the additional step of collapsing replicate trees using CONSENSE (also provided with PHYLIP). The raw data used for these analyses, including branch lengths and MEI allele frequencies, are provided in ED Table 4.

In a separate experiment, a hypothetical ancestor (ANS) that lacked all of the phase 3 MEIs (*i.e.*, had homozygous REF genotypes at all phase 3 MEI sites) was incorporated into the tree. As expected, the ancestor rooted the tree (bottom), and was very distant (with a branch length of 0.04109) from modern humans. Admixed individuals also were placed on the tree. The ASW and ACB populations clustered with the AFR clade (as expected due to high levels of AFR ancestry). The PEL and MXL populations clustered as outgroups of the EUR and SAS super-populations, likely reflecting higher levels of native ancestry, particularly in PEL individuals. The PUR and CLM populations clustered as outgroups of the EUR super-population, coinciding with a high proportion of European ancestry.

Branch-Specific Mobile Elements

MEIs were assigned to specific branches on the phylogenetic tree as follows. MEIs were inspected to determine whether each MEI was present in the twenty non-admixed populations described in the phylogenetic tree section above. Any MEI that was unique to a single population was assigned to that population. MEIs were assigned to an upper branch if they were present in 100% of constituent populations of a given branch. Branch-specific MEIs are listed in ED Table 4.

Admixture analysis

The admixture analysis outlined in **ED Figure 5** was performed using ADMIXTURE[66]. The cross validation test associated with ADMIXTURE indicated that K = 5 was optimal[67]. Genotypes were prepared for ADMIXTURE using VCFtools ver 0.1.12b[68] and PLINK ver 1.9[69]. ADMIXTURE was run with default parameters in accordance with the instruction manual provided at the ADMIXTURE download site (http://www.genetics.ucla.edu/software/admixture/ [2015]). **ED Figure 5** was sorted by the estimated majority ancestry for each superpopulation. Raw Q values generated by ADMIXTURE at K=5 for each individual are provided in ED Table 4. We note that while MEIs are homoplasy-free and thus useful as forensic markers, this is not an explicit requirement of ADMIXTURE.

Principle Components Analysis

To examine population structure using an orthogonal approach we performed Principle Components Analysis (PCA) using the same method as that used for deletions outlined above. The results largely corroborate population structure seen in deletions with PC1 separating AFR individuals from all other super-populations and PC2 separating EAS, SAS, and EUR individuals.. PUR and CLM individuals largely group within the EUR population with PEL and MXL populations intermediate to the EAS and EUR super-populations. As expected, African derived populations (ACB, ASW) largely clustered within the AFR super-population. PC3 and PC4 are similar to the population structure observed in deletions with the exception that there was a relative lack of separation of AMR individuals along PC4 with MEIs compared to deletions.

Inference of mutation rates and selective signatures from the site frequency spectrum (SFS)

The site frequency spectrum can be used to infer various population genetic parameters of different forms of genetic variation and to make inferences about selection. To estimate the mutation rate of various classes of SVs from our dataset we first estimated Waterson's θ:

$$\hat{\theta}_w = \frac{K}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

where K is the number of segregating sites and n is the number of chromosomes assessed. We then estimated the mutation rate as:

$$\hat{\theta}_w = 4N_e\mu$$

assuming an effective population size N_e of 10,000 (ref. [70]). Estimates of the mutation rate for various SV classes are reported in Table 6.1.6. We estimate a mutation rate of 0.113 deletions per haploid genome per generation, higher than previous estimates (*i.e.,* Conrad *et al.*[70]: 0.03; Kloosterman *et al.*[71]: 0.04), which can be explained by our increased power for detecting SVs <5 kbp. Indeed, when excluding calls <5 kbp we estimate a mutation rate of 0.037 consistent with earlier reports[70,71]. Additionally, our rate of novel *alu* insertion, 0.035 per haploid genome per generation, is very similar to that reported by Kloosterman *et al*.[71] (0.023).

Table 6.1.6: Estimates of Waterson's θ and the inferred mutation rate μ assuming an effective population size of 10,000.

N	θ	SV type	μ
1	0.10993955	DEL_HERV	2.75E-06
9	0.989455947	DEL_SVA	2.47E-05
56	6.156614781	DEL_LINE1	0.000153915
168	18.46984434	INS	0.000461746
786	86.41248603	INV	0.002160312
835	91.79952396	SVA	0.002294988
1238	136.1051625	DEL_ALU	0.003402629
3048	335.0957474	LINE1	0.008377394
5713	628.0846472	DUP	0.015702116
12748	1401.509379	ALU	0.035037734
32916	3618.770216	DEL (> 500bp)	0.090469255
40952	4502.244438	DEL	0.112556111

We also assessed the site frequency spectrum of bi-allelic SVs as a function of size, focusing only on deletions and duplications, variant classes with >1000 sites and sufficient spread in terms of SV sizes. We observe that as deletions increase in size, they become increasingly more rare, evidence of selection against events that are more likely to intersect functional elements (P<2.2e-16, linear model). For duplications we independently assessed events <15 kbp and >20 kbp, as different filtering parameters were used for calls below and above these size thresholds. We observed that SV size has little effect on the mean allele frequency of biallelic duplication events in either range (P=0.07 and P=0.382 respectively).

6.2.eQTL Analysis

Contributed by: Francesco Paolo Casale, Oliver Stegle

To ascertain the effect of SVs on gene expression, we considered 446 individuals from the GEUVADS consortium[72] that are overlapping with the 1000GP samples. We extracted the genotypes of SVs with VAF>1% from the 1000GP SV Analysis Group release set, which resulted in a set of 14,531 SVs considered for this analysis. We additionally obtained 11,514,964 SNPs and 1,296,114 indels (<50bp) from the 1000GP marker paper release set,

using this VAF cutoff. We employed a common 0/1/2 encoding (0: major hom, 1: het, 2: minor hom). For multi-allelic copy number variants (mCNVs), we considered the number of copies as a genetic feature. Allele frequencies for multi-allelic variants were calculated as mean copy number across genotyped individuals divided by the maximum copy number value for a given variant.

Gene expression levels were estimated using BitSeq[73] using the raw expression levels obtained from array express (http://www.ebi.ac.uk/ena/data/view/ERP001942 [2015]). Briefly, raw sequencing reads were aligned using Bowtie mapping to the GRCh37 reference build 69. Quantification of transcripts was pursued using BitSeq[73] version 0.4.3 with default settings. To map eQTLs, we used gene expression abundance estimates, which were obtained by averaging over multiple isoforms per gene. After removing lowly expressed genes, this resulted in 18,969 protein-coding genes from the autosomes, which we used for all of our eQTL analyses. Following the approach taken in[72], we used PEER [74] to estimate and account for hidden covariates and confounding factors. We estimated K=30 hidden factors using PEER with default parameter values. All subsequent analyses were performed on PEER adjusted expression data except for effect size estimates, which were obtained on unadjusted expression values.

cis eQTL mapping

We mapped *cis* eQTLs using a linear mixed model implemented in LIMIX[75], considering variants up to 1 Mbp up- or down-stream of each gene and jointly testing for associations using SVs and SNPs. In this analysis, population structure was accounted for using a random effect term and covariates were accounted for as fixed effects.

To correct for multiple testing within *cis* candidate regions of individual genes, we employed a permutation approach, where we permuted genetic markers relative to the covariates, random effect matrices and the phenotype (thereby retaining the relationship between the random effect and the expression trait). Both for permuted and non-permuted data we obtained the test statistics of the most associated variant within any *cis* region. Region-wise adjusted p-values were estimated by comparing the actual test statistics to the empirical null test statistics from 10,000 permutation experiments. To adjust for multiple testing across genes, we employed the Benjamini Hochberg procedure, estimating q-values for every gene. At FDR 10%, this resulted in a total number of 9,591 genes with an eQTL (egenes). A complete list of all discovered eQTLs is provided as ED Table 7.

Identification of LD-linked SV eQTLs

First, to identify strict SV-lead eQTLs, we considered each of the 9,591 egenes and tested whether the lead variant (minimal p-value) was an SV or a SNP/indel. This stringent criterion resulted in 54 strict SV-lead eQTLs (ED Table 8). This number may be an underestimate of the true number of SVs at MAF>1% affecting gene expression given expected systematic differences in genotyping accuracy between SNPs and SVs. We thus also considered the lead variants for each of the 9,591 egenes and tested for SVs in local LD

(within +/- 1 Mb window, r^2>0.5, a test only considering biallelic SVs). These LD association tests were implemented using PLINK[69], using the full panel of 2,504 individuals, yielding 166 additional egenes where an SV was in LD with a SNP lead eQTL. Taken together, this results in 220 eQTLs with evidence for an SV-implicated regulatory effect, 54 strict SV-lead eQTLs and 166 LD-linked SV eQTLs (ED Table 8). We found similar proportions of SV implicated eQTLs for secondary and tertiary associations (data not shown).

Enrichment analysis

We considered two alternative approaches to estimate the relative enrichment of SVs with a regulatory effect on gene expression compared to SNPs. First, we calculated the number of SV eQTLs relative to the number of SVs in *cis* candidate regions genome wide, comparing this to the relative proportion of SNP eQTLs versus the total number of SNP variants in the same regions. This basic enrichment score resulted in an up to 47-fold enrichment (mCNV; p<2.84E-39) of SVs when considering strict SV-lead eQTLs and in an up to 65-fold enrichment (mCNV; p< 8.01E-59) when considering LD-linked eQTLs; see ED Table 8). Statistical significance of the enrichments was assessed using a one-sided Fisher's exact test. To place the SV enrichment into context of shorter insertion and deletions (indels), we also considered eQTLs where one of 1,296,114 indels was the lead variant (similar to the primary analysis of the GEUVADIS data[72]). Notably, although indels explain a much larger number of eQTLs than SVs (1,339 versus 54 strict SV eQTLs and 220 LD-linked SV eQTLs), they were only marginally enriched compared to SNPs (1.4 fold, p<1.23E-30, ED Table 8). In the following analysis, we considered SNPs and indels together as these variant classes appeared to be comparable in terms of their potential to associate with gene expression.

We note that basic count-based enrichment may yield optimistic enrichments. This is because the number of effective SNP variants that are being tested within *cis* regions is substantially reduced due to strong local linkage patterns. We thus also considered a second enrichment strategy, where we compared the number of SVs that were in LD with a lead eQTL to a random expectation obtained from a sampling background model. To perform this enrichment analysis, we chose random locations in *cis* candidate regions and attempted to identify loci that match key properties of the 14,531 SVs that were considered in the *cis* eQTL analysis. These randomly drawn loci were approximately matched to real SVs, considering allele frequency (to the nearest 0.1 bin), haplotype length (within 50% size of each other) and distance to the TSS (within bins of 1000 bp). Here, haplotype length was estimated by the maximum distance of two variants close to the locus that have an r^2 value >= 0.80.

Using this strategy, we generated 100 random SV sets in *cis* candidate regions and assessed the proportion of these pseudo-SVs that were in LD with a *cis* eQTL. We used this random expectation to calculate the enrichment of genuine LD-linked SV-eQTLs (versus SNP/indel eQTLs), again resulting in a robust enrichment (ED Table 8) for this conservative enrichment testing approach.

In summary, we have carried out two complementary approaches to assess the enrichment of SV-implicated regulatory effect on gene expression compared to, and both analyses support a robust enrichment of SVs showing that SVs are likely to have an appreciable regulatory effect once occurring in the vicinity of genes.

SV-centric analysis of genic and coding SVs

In addition to the enrichment analysis jointly considering SVs and SNP/indel variants, we also performed an SV-centric analysis, through testing coding SVs in isolation for associations with gene expression levels (ED Table 8; multiple-testing adjustment using Benjamini Hochberg; FDR cutoff = 10%). When considering the full set of 559 coding SVs (at 1% VAF), this analysis yielded 89 coding SVs in associations with gene expression (*i.e.*, 20% of SVs showed association at the given FDR threshold). The most frequently associated variant types were deletions and CNVs. For example 46 out of 260 gene coding sequence affecting mCNVs (18%) and 35 out of a total of 159 coding deletions (27%) were eQTLs according to this analysis (ED Table 8). We reasoned that because of the relatively moderate size of the gEUVADIS cohort, power to detect genetic effects of coding SVs is limited, especially for genes that are lowly expressed as well as for SVs that are relatively rare. To better understand the limits of detection, we restricted this association analysis to SVs with increased allele frequency and to genes with increased levels of gene expression. As expected, the power to detect eQTL effects of common SVs in highly expressed genes were dramatically increased, e.g. resulting in the majority of highly common deletions (VAF>20%) intersecting the coding regions of highly expressed genes to be eQTLs (ED Table 8).

Taken together, this analysis suggests that a large proportion of coding SVs affect the transcriptome. We note that this analysis does not account for the possibility that genic or coding CNVs merely tag SNPs that are causing gene expression differences (however, for gene-coding SVs this likely affects only a small number of cases).

6.3. Evidence for RNA intermediates of sequences inserted at deletion breakpoints.

Contributed by: Alexej Abyzov, Nick Parrish, Eugene J. Gardner

Approximately 25% of assemble deletion breakpoints contained inserted sequences. These typically arise as errors during non-homologous end joining (NHEJ) or are copied from loci proximal to the deletion due to replication template switching[76]. Another potential source of such inserted sequences was recently described: RNA reverse-transcribed and integrated into the site of a double-strand break[77]. In addition, transposable elements have been shown to be integrated within deletions and at sites of double-strand breaks in cell culture[78,79] and in reference genomes[80-82]. We therefore examined the deletion breakpoints

described here for evidence that the inserted sequences could have been derived from an RNA intermediate.

Exclusion of introns from the inserted sequences provides the highest-confidence evidence of an RNA intermediate. No examples of intron exclusion were found in the inserted sequences described here (data not shown). Therefore we looked for evidence suggestive of an RNA intermediate (albeit not formally excluding a DNA source) namely 3' poly(dA) tails that may form as a consequence of template-primed reverse transcription of polyadenylated RNA. Allowing for a few non-A bases in the tail we identified 16 candidate SVs. We noted a preponderance of 3' poly(dA) tracts of at least 10 nucleotides, of which there were 12 examples (in contrast to only four 5' poly(dA) tracts). This differs significantly from the prediction, based on random sampling of DNA templates as predicted by template switching mechanisms, that one would observe on average 3.4 such tracts in either orientation in a set of 1,651 inserted sequences of length greater than 10 bp. Twelve 3' poly(dA) tracts represents a statistically significant increase from the predicted 3.4 (*p*-value of 2.2×10^{-4}). Based on statistical support for the notion that some of these inserted sequences could have been inserted through an RNA intermediate, we examined these poly(dA) containing inserted sequences in further detail. In one case (SV call id: UW_VH_1748), an RNA intermediate explains several features of the structural variant that template switching mechanisms cannot: the putative template DNA is on a distinct chromosome from the deletion and matched the inserted sequence with the exception of a poly(dA) of 36 bp. This variant can be parsimoniously explained if an mRNA transcribed from chromosome 2 was polyadenylated and inserted into a 410 bp deletion in chromosome four.

The remaining 15 poly(dA) containing inserted sequences were associated with the 3' termini of *Alu* elements from active subfamilies[83], consistent with their potential to be transcribed and mobilize. However, as *Alu* poly(dA) tracts are genome-encoded and are abundant as potential DNA templates in a template switching mechanism, we sought additional evidence to determine which of these sequence insertions, if any, were involved an RNA intermediate. We considered the degree of conservation of these inserted sequences to a given *Alu* subfamily, reasoning that highly mutated elements would be less likely to mobilize[83] and potentially have uniquely identifiable DNA templates. Four elements with divergence from their subfamily consensus at several nucleotides were identified. These four elements were likely inserted via a DNA template switching mechanism because sequences identical to the insertions were found on the same chromosome by BLAT or BLAST alignment. This left eleven 3' partial *Alu* elements within deletion breakpoints for which an RNA intermediate could not readily be excluded. By trimming poly-A/T tails and matching the remaining MI sequence with 50 kbp of deletion breakpoints, we identified potential template site for only 3 MIs, demonstrating that majority of these MI sequences are unlikely to be copied from local DNA templates. Next we examined the sequence at the deletion breakpoint for evidence of endonuclease cleavage consensus sequence. Allowing one deviation from the consensus, all eight remaining poly(dA) containing sequence insertions were within such sites. Thus these sequence insertions likely represent *Alu*-insertion associated deletions[84], albeit with a degree of 5' truncation.

We also sought to identify Non-Canonical *Alu* Insertions (NCAIs)[82] by looking at discordant read pairs that mapped to the human reference adjacent to the Phase III deletion calls. We examined all such discordant pairs to see if the unmapped mate aligned to an *Alu* consensus sequence[85]. After identifying candidate deletions, manual assembly was performed using Pacbio, Moleculo, or one of 30 high coverage genomes to confirm the presence of partial *Alu* sequences within the deletion breakpoints. Five sites were identified with this approach that were not identified by the assembly-based approach. Finally, we excluded DNA template switching or non-allelic homologous recombination using the same method outlined above, looking for an exact match in the human reference to the inserted sequence within 50 kbp. No such matches were found for any of the five insertions, thus leaving the only likely formation mechanism as NCAI (Alu insertions at NHEJ deletion – ED Table 13c).

6.4. Dispensable genes

Contributed by: John Huddleston, Evan Eichler

We identified 5,819 homozygous deletion genotypes from the complete set of structural variants (SVs) and annotated all events that completely deleted at least one exon (untranslated region, UTR or coding sequence, CDS) in the RefSeq gene annotations for GRCh37/hg19. For each SV that deleted at least one gene, we annotated the minimum residual variation intolerance score (RVIS) score of all affected genes[86]. Additionally, all homozygous deletions were annotated for their heterozygous and homozygous frequencies in each super population. With this approach, we identified 204 homozygous deletions affecting 240 genes. Based on the DAVID gene ontology classification[87], these genes were functionally enriched for immunoglobulin domains (Benjamini corrected p-value=1.0E-5) and glycoproteins (Benjamini corrected p-value=1.6E-3). Correspondingly, the mean of the minimum RVIS percentile of genes per homozygous deletion was 0.76, suggesting that these homozygously deleted genes are highly tolerant of mutation.

6.5. Overlap enrichment analysis of SVs versus genomic elements

Contributed by: Yan Zhang, Mark Gerstein

We performed permutation tests for several functional genomic elements (**ED Table 6.5.1.**, below) intersecting with SVs. We employed a "partial overlap statistic" and an "engulf overlap statistic" respectively in two series of tests, whereby the partial overlap statistic reflects the count of genomic elements (e.g. CDS) showing at least 1 bp overlap with SV intervals (e.g. deletions), and engulf overlap statistic reflects the count of genomic elements that are fully imbedded in at least one SV interval. In the permutation tests, the null distribution (random background) of the overlap measures is calculated from true genomic elements intersecting randomly shuffled SV locations. We generated 1,000 randomly

shuffled SV sets. Each shuffled set contains the same number of SVs, same proportion of SVs, and same length distribution as the real set. For deletions, we additionally generated 1000 randomly shuffled sets in each allele frequency bin of (0, 0.001], (0.001,0.01], and (0.01,1]. Taking heterogeneity of chromosomes into account, we required that shuffled SVs are still located on the same chromosome, and removed hg19 gap locations. BEDTools[88] was used for bed file operation and generating shuffled sets. The enrichment of genomic element-SV overlap is expressed as log2 fold change of the observed overlap statistic versus the mean of the null distribution. Positive (negative) log2 fold change indicates enriched (depleted) genomic element-SV overlap compared to random background. Each pair of genomic element type and SV type was tested individually. Empirical p-value were calculated, and reported to be significant if p-value <0.01. Error bars in the plots (**Figure 2ab, ED Figure 7**) reflect standard deviations of log2 fold changes in each permutation test.

In order to test whether CDS from a "low RVIS category" (e.g. RVIS<20) were more depleted of deletions than CDS with higher RVIS, we performed another set of permutation tests between each pair of RVIS categories. In each pairwise test, we pooled the CDS in both RVIS categories (e.g. "low" or "high" labeled). Then, we shuffle the RVIS labels of the CDS regions 1000 times to generate CDS pools with random RVIS labels. Overlap statistic (partial or engulf) of each RVIS category in the pool was calculated overlapping with deletions. This test statistic reflected the ratio of the overlap statistic between two RVIS categories (e.g., $\frac{\text{Partial overlap statistic of CDS with low RVIS label vs.deletions}}{\text{Partial overlap statistic of CDS with high RVIS label vs.deletions}}$). The observed test statistic was compared with the null distribution of test statistics calculated using the randomized pools. Empirical p-values were calculated for each pairwise test.

ED Table 6.5.1 – Genomic elements used in overlap enrichment analysis.

Index	Genomic elements	Description	Source
1	Gene	Annotated whole gene region	GENCODE v19(ref [89])
2	Gene low retroduplication	A subset of genes, with known paralogs and/or pseudogenes; Genes with the number of retroduplications (including paralogs and pseudogenes) in range [0, 10], ~95% in the subset	Gene and pseudogene annotation from GENCODE v19, gene-paralog pairs from Ensembl[90] gene-pseudogene pairs newly identified from PseudoPipe[91] for GENCODE v19 pseudogenes
3	Gene high retroduplication	A subset of genes, with known paralogs and/or pseudogenes; Gene with the number of retroduplications (including paralogs and pseudogenes) in range (10, 152], ~5% in the subset	
4	CDS	Annotated protein coding sequence region	GENCODE v19
5	CDS low RVIS	A subset of CDS with low Residual Variation Intolerance Score (RVIS)	CDS annotation from GENCODE v19, RVIS from literature[86]
6	CDS medium RVIS	A subset of CDS with medium RVIS	
7	CDS high RVIS	A subset of CDS with high RVIS	
8	Exon	Annotated exons in protein coding region	GENCODE v19
9	UTR	Annotated UTRs in protein coding region	GENCODE v19
10	Intron	Protein coding transcripts excluding exons	Processed from GENCODE v19
11	Pseudogene	Pseudogenes	GENCODE v19, requiring Type is transcript
12	Pseudogene processed	A subset of pseudogenes - processed pseudogenes	
13	Pseudogene unprocessed	A subset of pseudogenes - unprocessed pseudogenes	
14	Segmental duplication	Segmental duplication	(Eichler Lab)

15	lincRNA	Long, intervening noncoding RNAs	GENCODE v19
16	Ultraconserved	Ultraconserved regions across species	From literature[92,93]
17	Ultrasensitive nc	Ultrasensitive non-coding regions	From the study of 1000GP (Phase 1), Funseq[94]
18	ENCODE TF	ENCODE TF motif boundaries – more conserved regions in TF peak regions; overlapping intervals are merged.	Processed from data of ENCODE[95], Funseq[93,94]
19	TF peak	ENCODE TF peak region, union is taken for multiply reported peak regions	Processed from ENCODE TF peak data
20	piRNA Clusters	piRNA Clusters, filtered with RPKM ≥ 5	Processed from data published in literature[96]

Overlap enrichment analysis of SNPs versus genomic elements

We also performed the permutation tests for functional genomic elements intersecting with SNPs. The genomic elements used in this study are described above in the section "Overlap enrichment analysis of SVs versus genomic elements". We binned the SNVs into three allele frequency bins (0, 0.001], (0.001,0.01], and (0.01,1]. Similarly as in the SV analysis, we shuffled SNPs 1,000 times, while taking heterogeneity of chromosomes into account. We required the shuffled SNPs to be located on the same chromosome (removing assembly gaps). Partial overlap statistic was used, which is the count of genomic elements (e.g. CDS) that have at least 1 bp overlap with SNPs. The enrichment of genomic element-SNP overlap was expressed as log2 fold change of the observed overlap statistic versus the mean of the null distribution. Positive (negative) log2 fold change indicates enriched (depleted) genomic element-SNP overlap, compared to random background. Empirical p-value was calculated, and reported to be significant if p-value < 0.001. Error bars in the plots (**ED Figure 7**) indicate standard deviations of log2 fold changes in each permutation test.

6.6. Association of SVs with GWAS SNPs

Contributed by: Ryan Mills, Tobias Rausch, and Oliver Stegle

LD with GWAS SNPs

We sought to explore possible connections between our discovered SVs and previously reported SNPs that had been found to be associated with various phenotypes through genome-wide association studies (GWAS). To do this, we made use of the NHGRI Catalog of published GWAS (http://www.genome.gov/gwastudies/ [2015]) that describes 18,064 SNPs linked to a multitude of phenotypes. We cross-referenced this list with the set of genotyped phase 3 SNPs using their rs IDs, conservatively identifying 12,892 that were common to both sets. We then calculated the LD (r^2) between these SNPs and all SVs that had been reported within a 1 Mbp window using plink[69]. In this manner, we identified 136 SVs in strong LD (r^2>=0.8) with a GWAS SNP.

GWAS SNP Enrichment

We next explored whether we were observing a higher prevalence of GWAS SNPs in the flanking regions of SVs than we would expect from chance alone, an enrichment analysis controlled for VAF and haplotype size. To address potential biases in our enrichment

testing, we first removed redundancies between SNPs in high LD with each other and associated with the same phenotype. This resulted in a set of 12,495 GWAS SNPs. We focused on common SVs from our set with a minor allele frequency of >0.01 (n=9,188), of which roughly half (n=4,307) were present on a high-confidence haplotype block with definable length, which for the purpose of the analyses described in this chapter were defined as segments surrounding SVs having at least 1 SNP both upstream and downstream of the SV with r^2 >= 0.80 within 1 Mbp. (The flanking SNPs were used to define haplotype length.)

These data were summarized and stratified across different SV length bins (1 kbp-5 kbp-, 5 kbp-20 kbp, >20 kbp) and maximum r^2 values (0.4, 0.6, 0.8) and were then compared to our observed set. We observed a marked enrichment of GWAS SNPs in the flanks of larger (most pronounced for >20 kbp) SVs (when controlling for VAF and haplotype size). We additionally observed 1.75 fold enrichment for deletions <1 kbp, albeit not shown on the same **Figure 2** panel since these small deletions were genotyped using an alternative genotyping algorithm (*i.e.*, split-read based rather than read-depth based genotyping).

All custom software and relevant data sets utilized for this analysis can be found here: https://bitbucket.org/remills/1000gp_sv_phase3 [2015]

6.7. Personalized genomes analysis

Contributed by: Jieming Chen, Mark Gerstein, Oliver Stegle

Construction of personalized reference genomes

To study the effect of including SVs when constructing personalized reference genomes, we considered NA12878. Using the tool *vcf2diploid*[97] and additional custom scripts (http://alleledb.gersteinlab.org/docs/; citation: Chen J, Rozowsky J, Bedford J, Harmanci A, Abyzov A, Kong Y, Kitchen R, Regan L, Gerstein M. Allele-specific binding and expression: a uniform survey over many individuals and assays. *Manuscript submitted*.), we construct two alternative personalized reference genomes by incorporating phase 3 SNP (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/ALL.wgs.integrated_sv_map.20130502.svs.genotypes.vcf.gz [2015]) and SV variants (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20140708_previous_phase3/v4_vcfs/ [2015]) into the GRCh37 reference genome (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/ [2015]), only considering the 22 autosomes. First, we incorporated only SNPs and short indels, consisting of 3,548,153 SNVs and 554,853 indels. This reference will be referred to as the 'SNPs-genome'. In addition to this SNP/indel personalized genome, we considered a second reference, incorporating an additional set of 1,383 large SVs with breakpoint information (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/supporting/brea

kpoints/1KG_phase3_all_bkpts.v5.txt.gz [2015]) (termed 'SNPs+SVs-genome'). Because NA12878 is part of a CEU trio, most of the variants are phased into the paternal and maternal haplotypes, while those that are not, are randomly phased. In addition, we used the UCSC liftover tool (http://hgdownload.cse.ucsc.edu/admin/exe/ [2015]) to obtain matching personalized reference annotations based on hg19, considering a total of 54,839 genes and 280,213 exons in the GRCh37 reference (autosomes only). Using this approach a total of 280,179/280,181 (maternal/paternal) exons could be lifted on to the personalized genome reference using SNPs and 280,160/280,149 exons could be lifted to the personalized reference that included SNPs and SVs. Among these 280,123 exons were present in all reference genomes. We considered this set of consensus exons (out of 280,213 in the GRCh37 reference) to compare alternative genomes for read mapping (see below).

Read alignment and quantification

We used RNA-Sequencing reads for NA12878 obtained from Kilpinen *et al.*[98]. We then aligned the raw sequencing reads to alternative reference genomes, either considering the native GRCh37 reference, or separately aligning reads to the maternal and paternal genome of the two alternative personalized reference genomes (NA12878 SNPs- or SNPs+SVs-genome). Reads were aligned using STAR (Version 2.4.0h) with parameter settings:

1. Generation of the reference for alignment:
   STAR --runMode genomeGenerate --outTmpDir $tmp_dir --genomeDir $genome_dir --genomeFastaFiles $genome --runThreadN 6 --sjdbGTFfile $annotation --sjdbOverhang 100
2. Read alignment:
   STAR --genomeDir $genome_dir --outFilterMultimapNmax $multi_map_max --outFilterMismatchNmax 10 --alignIntronMax 500000 --alignMatesGapMax 1000000 --sjdbScore 2 --alignSJDBoverhangMin 1 --genomeLoad NoSharedMemory --outFilterScoreMinOverLread 0.33 --outSAMstrandField intronMotif --outSAMattributes NH HI NM MD AS XS --outSAMunmapped Within --outSAMtype BAM SortedByCoordinate --runThreadN $nthreads --readFilesIn $fasta_1 $fasta_2 --outFileNamePrefix $alignment_base
   #indexing
   samtools index $alignment_base/Aligned.sortedByCoord.out.bam

To count reads, we used custom scripts to quantify the number of reads mapping to genes and exons in hg19. For this analysis, only primary alignments were considered and both individual read pairs were counted separately. For the personalized genome approaches the union of unique reads (matched by read ID) mapped to the maternal or paternal alignment for a particular genomic feature were considered. All scripts used for the generation and the quantification of reads from personalized genomes are available on github (http://github.com/ostegle/SV1000gPersGenome [2015]).

Assessment approach

To assess the impact of alternative references, we considered the total number of reads mapped to 280,123 consensus exons that were present in all three reference annotations (autosomes only). Additionally, we considered the number of genomic features with at least 10, 100 or 1,000 reads mapped and the relative difference between alternative quantifications, requiring at least a difference of 10 reads and a one-fold difference (ED

Table 9).

Overall, a larger number of reads could be aligned to the SNPs+SVs-genome compared to the SNPs-genome. For example, 37,435,162 reads could be aligned to 280,123 consensus exons when using the GRCh37 reference. When considering the SNP-based reference, 37,702,108 reads mapped to the same set of exons (+266,946 reads) and 37,707,787 (+272,625 reads) mapped to the SNPs+SVs reference. The SNP-based reference resulted in 535 consensus exons with a profound change in expression compared to the GRCh37 reference (+/- 10 or more reads change, 1 fold change), whereas the SNPs+SVs reference resulted in 525 exons with a similar change in expression compared to the GRCh37 reference genome. The direct comparison between the SNP- and SNPs+SVs-genomes revealed 24 exons with a marked change in expression (+/- 10 or more reads change, 1 fold change), 18 of which were included in the set of 535 exons mentioned above.

The difference between the SV and SNP-based reference was even more dramatic when restricting the analysis to the 18 exons with a direct SV overlap. Six of these 18 exons were expressed (>10 reads) and of these, four exhibited substantial changes in expression (+/- 10 reads, 1 fold change) when using the SNPs+SVs-genome compared to the SNPs-genome (ED Table 9).

6.8. Features associated with SV clusters
Contributed by: John Huddleston, Evan Eichler

To investigate the relationship between SV hotspots and replication timing, we obtained Repli-seq data from the UCSC Genome Browser's UW Repli-seq track set. Specifically, we used the wavelet-smoothed signal track for a cell line from a normal individual who is also part of the 1,000 Genomes cohort (NA12878). The wavelet-smooth signal track provides a summary floating point value for each location in the genome where smaller values represent late cell cycle and larger values represent early cell cycle. We intersected SV clusters with this Repli-seq signal and calculated the weighted mean of the signal per cluster. We found no clear relationship across all clusters between distinct CNVRs per cluster and replication timing. Within the thirty GM12878 SV hotspots, 12 hotspots (40%) intersected with regions of late replication (wavelet signal < 20). The mean Repli-seq signal across all SV hotspots of 38 is slightly lower than expected for equivalently-sized regions in random genomic regions ($p = 0.012978$ with 1,000,000 permutations) suggesting that there is a bias for SV hotspots in regions of late replication.

SV clusters are predominately composed of deletions. Deletions are the most common SV in 2,260 of 3,163 clusters (71%) and 30 of 30 hotspots (100%). On average, deletions compose 66% of SVs in clusters and 63% of SVs in hotspots. Given that the overall proportion of deletions in the complete SV call set was 62%, the composition of SV clusters and hotspots is consistent with expectations.

In addition to checking for biases in replication timing and SV type, we compared SV clusters with previously described fragile sites[48] and genes affected by fragile sites (http://en.wikipedia.org/wiki/Chromosomal_fragile_site). To test fragile sites described in Table 1 of Mrasek *et al.*[48], we first mapped the cytoband locations for each site to corresponding coordinates in GRCh37/hg19 available from the UCSC Genome Browser. We intersected SV clusters with fragile site loci and assigned a binary value of 0 or 1 to each cluster based on the absence or presence of an overlap with a fragile site, respectively. Of 3,163 SV clusters, 1,960 (62%) overlapped with known fragile sites. This pattern represents a significant enrichment for fragile sites compared to equivalently-sized loci that were randomly distributed across the genome ($p = 0.001723$ with 1,000,000 permutations). Of the 30 SV hotspots (clusters containing >36 distinct CNVRs), 23 (77%) overlap with known fragile sites ($p = 0.035818$ with 1,000,000 permutations). Additionally, we found 11 of 16 (69%) previously described genes affected by fragile sites intersected with a SV cluster and 3 of 16 (19%) intersected with a SV hotspot.

6.9. Comparison of SVs to clinical genomics datasets
Contributed by: John Huddleston, Evan Eichler

We compared our dataset to a number of previously published clinical genomics studies to assess its utility to inform medical genetics. The studies assessed include: Yang *et al.*[99], The Deciphering Developmental Disorders [DDD] Study [100], Wright *et al.* [101], Boone *et al.*[102] Dittwald *et al.*[103] and Coe *et al.* [104]. We also compared de novo SVs from the DDD Study with all SVs from 1000 Genomes.

First, we assessed the overlap between disease-associated genes described in eTable 4 of Yang *et al.*[99]and homozygously deleted genes described in the 1000 Genomes samples. Of 318 distinct genes identified by Yang *et al.* 2014[99], only 2 (0.6%) were also identified in 1000 Genomes homozygous gene deletions specifically DEAF1 and TPM3. The deletion of DEAF1 in 1000 Genomes individuals deleted a 43bp exon in an alternate isoform (RNA accession: FJ985253.1). Similarly the deletion of TPM3 in the 1000 Genomes deleted the last exon of the smallest alternate isoform (RNA accession: NR_103460.1).

For the comparison of all SVs from the 1000 Genomes cohort with the 87 de novo CNVs reported by the Deciphering Developmental Disorders (DDD) Study[99], we used a standard 50% RO test and identified one deletion and two duplications shared between the two call sets. None of the 61 CNV-affected genes in the DDD Study overlapped with homozygous gene deletions from the 1000 Genomes. When we considered all genes in the DDD Study affected by SNVs whose predicted effects were greater than a missense (based on Ensembl's VEP ordered list: http://uswest.ensembl.org/info/genome/variation/predicted_data.html), only one gene, LEPREL1, overlapped with the set of 1000 Genomes homozygously deleted genes. The 1000 Genomes deletion of LEPREL1 deletes the first non-coding exon of one isoform (RNA accession: NM_001134418.1).

For the comparison of disease-associated genes described in Wright *et al.* 2015[101] and homozygously deleted genes in 1000 Genomes samples, we used the Developmental Disorders Genotype-to-Phenotype (*DDG2P*) database of 1,339 genes (Appendix 2, S1d) and the subset of 146 genes in that set that had an associated diagnosis among patients (Appendix 2, S2). Of the 1,339 genes, 5 (0.4%) overlapped with homozygously deleted genes in 1000 Genomes samples (*CFC1, DEAF1, GHR, HYAL1,* and *SCN11A*). There were no overlaps with the 146 genes associated with a patient diagnosis.

In our comparison between 1000 Genomes dispensable genes and other clinical studies, we found very few shared genes. Of the five homozygous gene deletions identified in Boone *et al.* 2013[102], only one (*LEPREL1*) was also present in the 1000 Genomes dispensable genes. Similarly, we only found 4 of 374 heterozygously deleted genes (1%) from Boone *et al.* 2013[102] in the dispensable genes list. For comparison with Dittwald *et al.* 2013[103], we selected the 232 disease-genes that overlapped directly-oriented paralogous low-copy repeats (DP-LCRs). Of these 232 genes, 5 (2%) overlap with dispensable genes from 1000 Genomes including *CFHR1, CFHR3, FCGR3B, GYPB,* and *SNRPN*. For comparison with Coe *et al.* 2014[104], we selected all gene deletions with Signature deletion p-value < 0.01 using the "Newest RefSeq Name" identifier for each gene. Between the 1,945 Signature deletions and dispensable genes from 1000 Genomes, there were 12 shared genes (0.6%) including *ADNP2, ATAD3B, ATAD3C, HBA1, HSBP1L1, IRAK2, KCNG2, LPAL2, PQLC1, RBFA, TBC1D21*, and *TXNL4A*.

Overall, the result that dispensable genes in the 1,000 Genomes cohort are rarely found in large disease cohorts is consistent with the notion that 1,000 Genomes samples represent to a great extent unaffected ("normal") individuals.

6.10. Evidence of Uniparental Disomy in 1000 Genomes Trio Families
Contributed by: Yu Kong and Adam Auton

Our SV callset includes a median number of homozygous deletions per individual of 557. We searched for instances of uniparental disomy (UPD) to investigate these homozygosity stretches (and additional stretches of homozygosity that may have escaped our SV calling strategy) in further detail. Specifically, to detect instances of UPD, we used SNP microarray data for trios generated on the OMNI and Affy platforms. In total, 407 trios were genotyped on the OMNI platform, 631 were genotyped on the Affy platform, with 404 trios having been genotyped on both. This data can be found on the 1000 Genomes FTP site in the following location: /vol1/ftp/release/20130502/supporting/hd_genotype_chip/

Using this data, we first identified all Mendelian errors within each trio. The median trio has 701 Mendelian errors on the OMNI platform (IQR: 332), and 616 on the Affy platform (IQR: 571). In the presence of UPD, we would expect to observe large numbers of Mendelian errors consistent with over transmission of alleles from one parent or the other. We therefore extracted errors consistent with UPD, and found the median trio to have 529 (robust $\sigma = 248$) and 418 (robust $\sigma = 259$) such errors on the OMNI and Affy platforms

respectively. We extracted 21 unique trios with more than 1,000 UPD-consistent Mendelian errors on one array or the other (Table).

Trio	Affy			Omni			Region of clustered UPD events	UPD derived from parent
	Total Mendel Errors	Maternal UPD-consistent errors	Paternal UPD-consistent errors	Total Mendel Errors	Maternal UPD-consistent errors	Paternal UPD-consistent errors		
HG01243_HG01241_HG01242	2225	726	358	1195	428	616	None	
HG01891_HG01890_HG01889	957	456	396	1176	455	588	None	
HG02222_HG02221_HG02220	700	208	427	1379	319	958	chr4 : 4.9Mb - 8.6Mb	Father
HG02492_HG02490_HG02491	13177	520	527				None	
HG02650_HG02648_HG02649	2789	1395	171				chr6 : 8.2Mb - 32.3Mb	Mother
HG02677_HG02675_HG02676	7431	591	573				None	
HG02776_HG02774_HG02775	6035	4295	238				chr12 : 43.1Mb - 134Mb	Mother
HG02871_HG02869_HG02870	1521	788	382				None	
HG03098_HG03096_HG03097	1156	583	540				None	
HG03110_HG03109_HG03108	1847	530	486				None	
HG03161_HG03160_HG03159	1708	732	440				None	
HG03269_HG03268_HG03267	1189	469	534				None	
HG03374_HG03373_HG03372	8170	519	483				None	
HG03453_HG03451_HG03452	1147	587	515				None	
NA12329_NA06984_NA06989	1019	590	101	2544	1407	266	chr18 : 67.0Mb - 78.0Mb	Mother
NA12865_NA12874_NA12875	1081	610	110	2049	1131	135	chr1 : 238Mb - 249Mb	Mother
NA18497_NA18498_NA18499	1127	571	456	1315	680	440	chr1 : 22.3Mb - 28.5Mb	Mother
NA18518_NA18519_NA18520	1197	425	724	1293	453	649	chr11 : 0Mb - 27.8Mb (Affy Only)	Father
NA19208_NA19207_NA19206	1282	384	789	1240	365	536	chr9 (Affy Only)	Father
NA19742_NA19741_NA19740	94991	42333	118	170397	74398	205	Whole Genome	Mother
NA19918_NA19916_NA19917	1189	369	716	1521	410	989	chr17 : 0Mb - 8.3Mb	Father

Table 6.10.1: Trios with > 1000 UPD-consistent Mendel errors.

We visually inspected these 21 trios for clusters of UPD-consistent errors along the genome. Of the 21, 11 showed no clear evidence of clustering of UPD transmission events within localized regions of the genome. A further two trios showed potential UPD events, but only on the Affy array, potentially indicative of cell-line artifacts. Finally, one trio showed evidence of maternal UPD across the whole genome, potentially indicative of a sample mix-up, or other artifact.

The remaining 7 trios contained localized regions with clusters of UPD transmissions from one parent or the other. These regions averaged 22.1Mb in size, ranging from 3.7Mb to 90.9Mb. Only 1 in 240 (0.5%) of genes in our homozygous gene deletion knockout list coincided with these seven potential UPD stretches. Furthermore, there was no evidence of preferential UPD from one parent or the other, with 5 UPD events transmitted from the mother, and 2 from the father.

Given the sample size of 634 trios, we estimate the rate of UPD to be of the order of 1%. This is considerably higher than the estimates of ~0.02% presented in the literature[105]. And while our resolution for mapping these events is higher compared to cytogenetics studies, we note that our estimate is likely an overestimate of the true incidence of UPD for a couple of reasons. First, the DNA from our study is derived from cell lines, which may contain somatic genomic alterations that have occurred since transformation. Indeed, we note that 4 of the 10 samples containing evidence of putative UPD regions are derived from the CEU and YRI populations, which represent the oldest cell lines within the 1,000 Genomes Project collection, and that two of the putative UPD regions are only detected using one microarray platform. Second, all but one of our detected UPD regions are less than 25Mb in size, and do not cover whole chromosomes as might be expected under many UPD casual mechanisms[106]. As such, many of our detected regions may be the result of localized somatic genome alterations leading to mis-identification of UPD.

6.11. Long Regions of Homozygosity 1000 Genomes Individuals
Contributed by: Adam Auton, John Huddleston, Peter Sudmant

Runs of homozygosity (*i.e.*, extended stretches of a genome without of heterozygous variants) occur within an individual when both homologous chromosomes share a recent common ancestor. To quantify the expected levels of homozygosity in the 1,000 Genomes samples, we applied a HMM procedure[107] to identify long runs of homozygosity (LROH) within the 1000 Genomes samples. Using the Phase 3 callset, we identified LROH across the autosomes for each individual using VCFtools version 0.1.13[68] with the following command:

vcftools --gzvcf ALL.chr*XXX*.phase3_shapeit2_mvncall_integrated_v4.20130502.genotypes.vcf.gz --LROH --chr *XXX* --keep *population_file.txt*

where *XXX* represents the chromosome number, and *population_file.txt* represents a file that lists all of the individuals within a given population. The locations of the identified LROH were converted from physical distances to genetic distances using the sex-averaged map from Campbell *et al*. [108], and LROH less than were 1cM discarded.

On average, samples within the 1000 Genomes contain 23.0 cM of sequence within LROHs, and representing approximately 0.7% of the genome (median: 8.8 cM; standard deviation: 46.7 cM). However, there is considerable variation between individuals, with 109 individuals (~4%) having no LROH longer than 1 cM, and 25 individuals (~1%) having more than 250 cM of their genome contained within LROHs. In addition, the extent of LROH varies between populations, with the lowest levels of LROH found within the ASW and ACB samples (median: 2.1 cM and 3.4 cM respectively), and the highest within the STU and PJL samples (median: 41.7 cM and 29.7 cM respectively).

We additionally searched for individuals harboring long autozygos tracts, homozygous stretches of 5 Mbp or greater[109]. We identified 447 individuals harboring at least a single autozygos tract. Particularly large amounts of autozyosity were observed in PJL, STU, ITU and CLM populations.

Furthermore, we searched for an association of LROH with complex SVs. Of the 447 samples with a LROH >=5 Mbp, 33 (7%) had a complex SV that mapped within the LROH. This proportion is less than expected compared to a mean of 115 samples from a null distribution created by counting samples with shared complex SVs in randomly placed LROH regions >=5 Mbp (empirical $p < 0.00001$ after 100,000 iterations). Thus, compared to previously presented data from clinical samples (Carvalho *et al*.[110]) our data shows that an association between complex SVs and large homozygous regions is uncommon in healthy SV carriers.

References in the Supplementary Material

1 Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6**, 677-681, doi:nmeth.1363 [pii] 10.1038/nmeth.1363 (2009).

2 Handsaker, R. E. *et al.* Large multiallelic copy number variations in humans. *Nature genetics*, doi:10.1038/ng.3200 (2015).

3 Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333-i339, doi:10.1093/bioinformatics/bts378 bts378 [pii] (2012).

4 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 btp324 [pii] (2009).

5 Hormozdiari, F., Alkan, C., Eichler, E. E. & Sahinalp, S. C. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res* **19**, 1270-1278, doi:10.1101/gr.088633.108 (2009).

6 Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **21**, 974-984, doi:gr.114876.110 [pii] 10.1101/gr.114876.110 (2011).

7 Hach, F. *et al.* mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat Methods* **7**, 576-577, doi:10.1038/nmeth0810-576 (2010).

8 Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641-646, doi:10.1126/science.1197005 (2010).

9 Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641-646, doi:330/6004/641 [pii] 10.1126/science.1197005 (2010).

10 Sudmant, P. H. *et al.* Evolution and diversity of copy number variation in the great ape lineage. *Genome Res* **23**, 1373-1382, doi:10.1101/gr.158543.113 (2013).

11 Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-226, doi:10.1126/science.1224344 (2012).

12 Witkin, A. P. in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84.* 150-153.

13 Mills, R. E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59-65, doi:10.1038/nature09708 (2011).

14 Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet* **43**, 269-276, doi:ng.768 [pii] 10.1038/ng.768 (2011).

15 Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871, doi:10.1093/bioinformatics/btp394 (2009).

16 Gardner, E. J. & Devine, S. E. MELT: Mobile Element Location Tool. *In Preparation* (2014).

17 Dombroski, B. A., Scott, A. F. & Kazazian, H. H., Jr. Two additional potential retrotransposons isolated from a human L1 subfamily that contains an active retrotransposable element. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 6513-6517 (1993).

18 Stewart, C. *et al.* A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet* **7**, e1002236, doi:10.1371/journal.pgen.1002236 PGENETICS-D-10-00611 [pii] (2011).
19 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).
20 Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-2993, doi:10.1093/bioinformatics/btr509 btr509 [pii] (2011).
21 Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res* **12**, 996-1006, doi:10.1101/gr.229102. Article published online before print in May 2002 (2002).
22 Dayama, G., Emery, S. B., Kidd, J. M. & Mills, R. E. The genomic landscape of polymorphic human nuclear mitochondrial insertions. *Nucleic Acids Res*, doi:10.1093/nar/gku1038 (2014).
23 Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-181, doi:10.1038/nmeth.1785 (2012).
24 Menelaou, A. & Marchini, J. Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. *Bioinformatics* **29**, 84-91, doi:10.1093/bioinformatics/bts632 (2013).
25 Chen, K. *et al.* TIGRA: a targeted iterative graph routing assembler for breakpoint assembly. *Genome Res* **24**, 310-317, doi:10.1101/gr.162883.113 (2014).
26 Michaelson, J. J. & Sebat, J. forestSV: structural variant discovery through statistical learning. *Nat Methods* **9**, 819-821, doi:10.1038/nmeth.2085 (2012).
27 Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-829, doi:10.1101/gr.074492.107 gr.074492.107 [pii] (2008).
28 Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664, doi:10.1101/gr.229202. Article published online before March 2002 (2002).
29 A. Ummat, R. S., M. Pendleton, A. Pang, O. Franzen, T. Rausch, W. Stedman, T. Anantharman, A. Hastie, H. Dai, H. Cao, A. Cohain, G. Deikus,R. Durret, S. Blanchard, R. Altman, C.S. Chin, Yan Guo, E. Paxinos, J. Korbel, R.B. Darnell, W.R. McCombie, C.E. Mason, P.-Y. Kwok, E.E. Schadt, A. Bashir. in preparation.
30 Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238, doi:10.1186/1471-2105-13-238 (2012).
31 Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**, 563-569, doi:10.1038/nmeth.2474 (2013).
32 Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276-277 (2000).
33 Abyzov, A. & Gerstein, M. AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics* **27**, 595-603, doi:10.1093/bioinformatics/btq713 (2011).
34 Delcher, A. L., Salzberg, S. L. & Phillippy, A. M. Using MUMmer to identify similar regions in large sequence sets. *Curr Protoc Bioinformatics* **Chapter 10**, Unit 10 13, doi:10.1002/0471250953.bi1003s00 (2003).
35 Conrad, D. F. *et al.* Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nat Genet* **42**, 385-391, doi:ng.564 [pii] 10.1038/ng.564 (2010).

36 Mills, R. E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59-65, doi:10.1038/nature09708 (2011).
37 Lam, H. Y. *et al.* Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat Biotechnol* **28**, 47-55, doi:nbt.1600 [pii]
10.1038/nbt.1600 (2010).
38 Abyzov, A. *et al.* Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nat Commun* **6**, 7256, doi:10.1038/ncomms8256 (2015).
39 Koressaar, T. & Remm, M. Enhancements and modifications of primer design program Primer3. *Bioinformatics* **23**, 1289-1291, doi:10.1093/bioinformatics/btm091 (2007).
40 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).
41 Blankenberg, D. *et al.* Manipulation of FASTQ data with Galaxy. *Bioinformatics* **26**, 1783-1785, doi:10.1093/bioinformatics/btq281 (2010).
42 Giardine, B. *et al.* Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* **15**, 1451-1455, doi:10.1101/gr.4086505 (2005).
43 Smit, A., Hubley, R. & Green, P. *RepeatMasker Open-3.0*, <www.repeatmasker.org> (1996-2010).
44 You, F. M. *et al.* BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics* **9**, 253, doi:10.1186/1471-2105-9-253 (2008).
45 Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**, 365-386 (2000).
46 Hall, T. A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* **41**, 95-98 (1999).
47 Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12, doi:10.1186/gb-2004-5-2-r12 (2004).
48 Mrasek, K. *et al.* Global screening and extended nomenclature for 230 aphidicolin-inducible fragile sites, including 61 yet unreported ones. *Int J Oncol* **36**, 929-940 (2010).
49 Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64, doi:10.1038/nature06862 (2008).
50 Chaisson, M. J. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608-611, doi:10.1038/nature13907 (2015).
51 Parsons, J. D. Miropeats: graphical DNA sequence comparisons. *Comput Appl Biosci* **11**, 615-619 (1995).
52 Steinberg, K. M. *et al.* Single haplotype assembly of the human genome from a hydatidiform mole. *Genome Res* **24**, 2066-2076, doi:10.1101/gr.180893.114 (2014).
53 Korn, J. M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* **40**, 1253-1260, doi:10.1038/ng.237 (2008).
54 Altshuler, D. M. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).
55 MacDonald, J. R., Ziman, R., Yuen, R. K., Feuk, L. & Scherer, S. W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res* **42**, D986-992, doi:10.1093/nar/gkt958 (2014).
56 Stewart, C. *et al.* A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet* **7**, e1002236, doi:10.1371/journal.pgen.1002236 (2011).

57 Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30**, 97-101, doi:10.1038/ng786 (2002).
58 Do, R. *et al.* No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nat Genet* **47**, 126-131, doi:10.1038/ng.3186 (2015).
59 Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nat Genet* **46**, 220-224, doi:10.1038/ng.2896 (2014).
60 Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190, doi:10.1371/journal.pgen.0020190 (2006).
61 Jakobsson, M. *et al.* Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**, 998-1003, doi:10.1038/nature06742 (2008).
62 Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444-454, doi:nature05329 [pii] 10.1038/nature05329 (2006).
63 Felsenstein, J. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 164-166 (1989).
64 Cavalli-Sforza, L. L. & Edwards, A. W. Phylogenetic analysis. Models and estimation procedures. *Am J Hum Genet* **19**, 233-257 (1967).
65 Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406-425 (1987).
66 Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655-1664, doi:10.1101/gr.094052.109 (2009).
67 Alexander, D. H. & Lange, K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12**, 246, doi:10.1186/1471-2105-12-246 (2011).
68 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158, doi:10.1093/bioinformatics/btr330 (2011).
69 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575, doi:10.1086/519795 (2007).
70 Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704-712, doi:10.1038/nature08516 (2010).
71 Kloosterman, W. P. *et al.* Characteristics of de novo structural changes in the human genome. *Genome Res*, doi:10.1101/gr.185041.114 (2015).
72 Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506-511, doi:10.1038/nature12531 (2013).
73 Glaus, P., Honkela, A. & Rattray, M. Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics* **28**, 1721-1728, doi:10.1093/bioinformatics/bts260 (2012).
74 Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* **7**, 500-507, doi:10.1038/nprot.2011.457 (2012).
75 Lippert, C., Casale, F. P., Rakitsch, B. & Stegle, O. LIMIX: genetic analysis of multiple traits. *bioRxiv doi: 10.1101/003905* (2014).
76 Zhang, F. *et al.* The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat Genet* **41**, 849-853, doi:10.1038/ng.399 (2009).

77 Onozawa, M. *et al.* Repair of DNA double-strand breaks by templated nucleotide sequence insertions derived from distant regions of the genome. *Proc Natl Acad Sci U S A* **111**, 7729-7734, doi:10.1073/pnas.1321889111 (2014).
78 Gilbert, N., Lutz-Prigge, S. & Moran, J. V. Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110**, 315-325 (2002).
79 Morrish, T. A. *et al.* DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nat Genet* **31**, 159-165, doi:10.1038/ng898 (2002).
80 Sen, S. K., Huang, C. T., Han, K. & Batzer, M. A. Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome. *Nucleic Acids Res* **35**, 3741-3751, doi:10.1093/nar/gkm317 (2007).
81 Han, K. *et al.* Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. *Nucleic Acids Res* **33**, 4040-4052, doi:10.1093/nar/gki718 (2005).
82 Srikanta, D. *et al.* An alternative pathway for Alu retrotransposition suggests a role in DNA double-strand break repair. *Genomics* **93**, 205-212, doi:10.1016/j.ygeno.2008.09.016 (2009).
83 Bennett, E. A. *et al.* Active Alu retrotransposons in the human genome. *Genome Res* **18**, 1875-1883, doi:10.1101/gr.081737.108 (2008).
84 Callinan, P. A. *et al.* Alu retrotransposition-mediated deletion. *J Mol Biol* **348**, 791-800, doi:10.1016/j.jmb.2005.02.043 (2005).
85 E. J. Gardner, S. E. D. in preparation.
86 Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* **9**, e1003709, doi:10.1371/journal.pgen.1003709 (2013).
87 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57, doi:10.1038/nprot.2008.211 (2009).
88 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:btq033 [pii] 10.1093/bioinformatics/btq033 (2010).
89 Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-1774, doi:10.1101/gr.135350.111 22/9/1760 [pii] (2012).
90 Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res* **42**, D749-755, doi:10.1093/nar/gkt1196 (2014).
91 Zhang, Z. *et al.* PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics* **22**, 1437-1439, doi:10.1093/bioinformatics/btl116 (2006).
92 Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321-1325, doi:10.1126/science.1098119 (2004).
93 Fu, Y. *et al.* FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* **15**, 480, doi:10.1186/s13059-014-0480-5 (2014).
94 Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587, doi:10.1126/science.1235587 (2013).
95 Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 nature11247 [pii] (2012).
96 Ha, H. *et al.* A comprehensive analysis of piRNAs from adult human testis and their relationship with genes and mobile elements. *BMC Genomics* **15**, 545, doi:10.1186/1471-2164-15-545 (2014).

97 Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* **7**, 522, doi:10.1038/msb.2011.54 (2011).
98 Kilpinen, H. *et al.* Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* **342**, 744-747, doi:10.1126/science.1242463 (2013).
99 Yang, Y. *et al.* Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* **312**, 1870-1879, doi:10.1001/jama.2014.14601 (2014).
100 Deciphering Developmental Disorders, S. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223-228, doi:10.1038/nature14135 (2015).
101 Wright, C. F. *et al.* Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* **385**, 1305-1314, doi:10.1016/S0140-6736(14)61705-0 (2015).
102 Boone, P. M. *et al.* Deletions of recessive disease genes: CNV contribution to carrier states and disease-causing alleles. *Genome Res* **23**, 1383-1394, doi:10.1101/gr.156075.113 (2013).
103 Dittwald, P. *et al.* Inverted low-copy repeats and genome instability--a genome-wide analysis. *Hum Mutat* **34**, 210-220, doi:10.1002/humu.22217 (2013).
104 Coe, B. P. *et al.* Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat Genet* **46**, 1063-1071, doi:10.1038/ng.3092 (2014).
105 Liehr, T. Cytogenetic contribution to uniparental disomy (UPD). *Mol Cytogenet* **3**, 8, doi:10.1186/1755-8166-3-8 (2010).
106 Robinson, W. P. Mechanisms leading to uniparental disomy and their clinical consequences. *Bioessays* **22**, 452-459, doi:10.1002/(SICI)1521-1878(200005)22:5<452::AID-BIES7>3.0.CO;2-K (2000).
107 Auton, A. *et al.* Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Res* **19**, 795-803, doi:10.1101/gr.088898.108 (2009).
108 Campbell, C. L., Furlotte, N. A., Eriksson, N., Hinds, D. & Auton, A. Escape from crossover interference increases with maternal age. *Nat Commun* **6**, 6260, doi:10.1038/ncomms7260 (2015).
109 Campbell, C. D. *et al.* Estimating the human mutation rate using autozygosity in a founder population. *Nat Genet* **44**, 1277-1281, doi:10.1038/ng.2418 (2012).
110 Carvalho, C. M. *et al.* Absence of heterozygosity due to template switching during replicative rearrangements. *Am J Hum Genet* **96**, 555-564, doi:10.1016/j.ajhg.2015.01.021 (2015).

Author list - The 1000 Genomes Consortium

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

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

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

Analysis Group: Affymetrix Jeanette P. Schmidt (Principal Investigator)[24], Christopher J. Davies[24], Jeremy Gollub[24], Teresa Webster[24], Brant Wong[24], Yiping Zhan[24], **Albert Einstein College of Medicine** Adam Auton (Principal Investigator)[1], Christopher L. Campbell[1], Yu Kong[1], Anthony Marcketta[1] **Baylor College of Medicine** Richard A. Gibbs (Principal Investigator)[14], Fuli Yu (Project Leader)[14], Lilian Antunes[14], Matthew Bainbridge[14], Donna Muzny[14], Aniko Sabo[14], Zhuoyi Huang[14] **BGI-Shenzhen** Jun Wang (Principal Investigator)[26-30], Lachlan J.M. Coin[26], Lin Fang[26,27], Xiaosen Guo[26], Xin Jin[26], Guoqing Li[26], Qibin Li[26], Yingrui Li[26], Zhenyu Li[26], Haoxiang Lin[26], Binghang Liu[26], Ruibang Luo[26], Haojing Shao[26], Yinlong Xie[26], Chen Ye[26], Chang Yu[26], Fan Zhang[26], Hancheng Zheng[26], Hongmei Zhu[26], **Bilkent University** Can Alkan[36], Elif Dal[36], Fatma Kahveci[36], **Boston College** Gabor T. Marth (Principal Investigator)[23], Erik P. Garrison (Project Lead)[4], Deniz Kural[37], Wan-Ping Lee[37], Wen Fung Leong[38], Michael Stromberg[39], Alistair N. Ward[23], Jiantao Wu[39], Mengyao Zhang[40], **Broad Institute of MIT and Harvard** Mark J. Daly (Principal Investigator)[13], Mark A. DePristo (Project Leader)[41], Robert E. Handsaker (Project Leader)[13,40], David M. Altshuler[3], Eric Banks[13], Gaurav Bhatia[13], Guillermo del Angel[13], Stacey B. Gabriel[13], Giulio Genovese[13], Namrata Gupta[13], Heng Li[13], Seva Kashin[13,40], Eric S. Lander[13], Steven A. McCarroll[13,40], James C. Nemesh[13], Ryan E. Poplin[13], **Cold Spring Harbor Laboratory** Seungtai C. Yoon (Principal Investigator)[42], Jayon Lihm[42], Vladimir Makarov[43], **Cornell University** Andrew G. Clark (Principal Investigator)[7], Srikanth Gottipati[44], Alon Keinan[7], Juan L. Rodriguez-Flores[45], **European Molecular Biology Laboratory** Jan O. Korbel (Principal Investigator)[12,17], Tobias Rausch(Project Leader)[17,46], Markus H. Fritz[46], Adrian M. Stütz[17], **European Molecular Biology Laboratory, European Bioinformatics Institute** Paul Flicek (Principal Investigator)[12], Kathryn Beal[12], Laura Clarke[12], Avik Datta[12], Javier Herrero[47], William M. McLaren[12], Graham R.S. Ritchie[12], Richard E. Smith[12], Daniel Zerbino[12], Xiangqun Zheng-Bradley[12], **Harvard University** Pardis C. Sabeti (Principal Investigator)[13,48], Ilya Shlyakhter[13,48], Stephen F. Schaffner[13,48], Joseph Vitti[13,49], **Human Gene Mutation Database** David N. Cooper (Principal Investigator)[50], Edward V. Ball[50], Peter D. Stenson[50], **Illumina** David R. Bentley (Principal Investigator)[5], Bret Barnes[39], Markus Bauer[5], R. Keira Cheetham[5], Anthony Cox[5], Michael Eberle[5], Sean Humphray[5], Scott Kahn[39], Lisa Murray[5], John Peden[5], Richard Shaw[5], **Icahn School of Medicine at Mount Sinai** Eimear E. Kenny (Principal Investigator)[51]**, Louisiana State University** Mark A. Batzer (Principal Investigator)[52], Miriam K. Konkel[52], Jerilyn A. Walker[52], **Massachusetts General Hospital** Daniel G. MacArthur (Principal Investigator)[53], Monkol Lek[53], **Max Planck Institute for Molecular Genetics** Ralf Sudbrak (Project Leader)[32], Vyacheslav S. Amstislavskiy[20], Ralf Herwig[20], **McDonnell Genome Institute at Washington University** Elaine R. Mardis (Co-Principal Investigator)[22], Li Ding[22], Daniel C. Koboldt[22], David Larson[22], Kai Ye[22], **McGill University** Simon Gravel[54], **National Eye Institute, NIH** Anand Swaroop[55], Emily Chew[55], **New York Genome Center** Tuuli Lappalainen (Principal Investigator)[56,57], Yaniv Erlich (Principal Investigator)[56,58], Melissa Gymrek[13,56,59,60], Thomas Frederick Willems[61], **Ontario Institute for Cancer Research** Jared T. Simpson[62], **Pennsylvania State University** Mark D. Shriver (Principal Investigator)[63], **Rutgers Cancer Institute of New Jersey** Jeffrey A. Rosenfeld (Principal Investigator)[64], **Stanford University** Carlos D. Bustamante (Principal Investigator)[65], Stephen B. Montgomery (Principal Investigator)[66], Francisco M. De La Vega (Principal Investigator)[65], Jake K. Byrnes[67], Andrew W. Carroll[68], Marianne K. DeGorter[66], Phil Lacroute[65], Brian K. Maples[65], Alicia R. Martin[65], Andres Moreno-Estrada[65,69], Suyash S. Shringarpure[65], Fouad Zakharia[65], **Tel-Aviv University** Eran

Halperin (Principal Investigator)[70-72], Yael Baran[70], **The Jackson Laboratory for Genomic Medicine** Charles Lee (Principal Investigator)[18,19], Eliza Cerveira[18], Jaeho Hwang[18], Ankit Malhotra (Co-Project Lead)[18], Dariusz Plewczynski[18], Kamen Radew[18], Mallory Romanovitch[18], Chengsheng Zhang(Co-Project Lead)[18], **Thermo Fisher Scientific** Fiona C.L. Hyland[73], **Translational Genomics Research Institute** David W. Craig (Principal Investigator)[74], Alexis Christoforides[74], Nils Homer[75], Tyler Izatt[74], Ahmet A. Kurdoglu[74], Shripad A. Sinari[74], Kevin Squire[76], **US National Institutes of Health** Stephen T. Sherry (Principal Investigator)[25], Chunlin Xiao[25], **University of California, San Diego** Jonathan Sebat (Principal Investigator)[77,78], Danny Antaki[77], Madhusudan Gujral[77], Amina Noor[77], Kenny Ye[79], **University of California, San Francisco** Esteban G. Burchard (Principal Investigator)[80], Ryan D. Hernandez (Principal Investigator)[80-82], Christopher R. Gignoux[80], **University of California, Santa Cruz** David Haussler (Principal Investigator)[83,84], Sol J. Katzman[83], W. James Kent[83], **University of Chicago** Bryan Howie[85], **University College London** Andres Ruiz-Linares (Principal Investigator)[86], **University of Geneva** Emmanouil T. Dermitzakis (Principal Investigator)[87-89], **University of Maryland School of Medicine** Scott E. Devine (Principal Investigator)[90], **University of Michigan** Gonçalo R. Abecasis (Principal Investigator) (Co-Chair)[2], Hyun Min Kang (Project Leader)[2], Jeffrey M. Kidd (Principal Investigator)[91,92], Tom Blackwell[2], Sean Caron[2], Wei Chen[93], Sarah Emery[92], Lars Fritsche[2], Christian Fuchsberger[2], Goo Jun[2,94], Bingshan Li[95], Robert Lyons[96], Chris Scheller[2], Carlo Sidore[2,97,98], Shiya Song[91], Elzbieta Sliwerska[92], Daniel Taliun[2], Adrian Tan[2], Ryan Welch[2], Mary Kate Wing[2], Xiaowei Zhan[99] **University of Montréal** Philip Awadalla (Principal Investigator)[62,100], Alan Hodgkinson[100], **University of North Carolina at Chapel Hill** Yun Li[101], **University of North Carolina at Charlotte** Xinghua Shi (Principal Investigator)[102], Andrew Quitadamo[102], **University of Oxford** Gerton Lunter (Principal Investigator)[8], Gil A. McVean (Principal Investigator) (Co-Chair)[8,9], Jonathan L. Marchini (Principal Investigator)[8,9], Simon Myers (Principal Investigator)[8,9], Claire Churchhouse[9], Olivier Delaneau[9,87], Anjali Gupta-Hinch[8], Warren Kretzschmar[8], Zamin Iqbal[8], Iain Mathieson[8], Androniki Menelaou[9,103], Andy Rimmer[87], Dionysia K. Xifara[8,9], **University of Puerto Rico** Taras K. Oleksyk (Principal Investigator)[104], **University of Texas Health Sciences Center at Houston** Yunxin Fu (Principal Investigator)[94], Xiaoming Liu[94], Momiao Xiong[94], **University of Utah** Lynn Jorde (Principal Investigator)[105], David Witherspoon[105], Jinchuan Xing[106], **University of Washington** Evan E. Eichler (Principal Investigator)[10,11], Brian L. Browning (Principal Investigator)[107], Sharon R. Browning (Principal Investigator)[108], Fereydoun Hormozdiari[10], Peter H. Sudmant[10], **Weill Cornell Medical College**, Ekta Khurana (Principal Investigator)[109], **Wellcome Trust Sanger Institute** Richard M. Durbin (Principal Investigator)[4], Matthew E. Hurles (Principal Investigator)[4], Chris Tyler-Smith (Principal Investigator)[4], Cornelis A. Albers[110,111], Qasim Ayub[4], Senduran Balasubramaniam[4], Yuan Chen[4], Vincenza Colonna[4,112], Petr Danecek[4], Luke Jostins[8], Thomas M. Keane[4], Shane McCarthy[4], Klaudia Walter[4], Yali Xue[4], **Yale University** Mark B. Gerstein (Principal Investigator)[113-115], Alexej Abyzov[116], Suganthi Balasubramanian[115], Jieming Chen[113], Declan Clarke[117], Yao Fu[113], Arif O. Harmanci[113], Mike Jin[115], Donghoon Lee[113], Jeremy Liu[115], Xinmeng Jasmine Mu[13,113], Jing Zhang[113,115], Yan Zhang[113,115]

Structural Variation Group: BGI-Shenzhen Yingrui Li[26], Ruibang Luo[26], Hongmei Zhu[26], **Bilkent University** Can Alkan[36], Elif Dal[36], Fatma Kahveci[36], **Boston College** Gabor T. Marth (Principal Investigator)[23], Erik P. Garrison[4], Deniz Kural[37], Wan-Ping Lee[37], Alistair N.

Ward[23], Jiantao Wu[23], Mengyao Zhang[23], **Broad Institute of MIT and Harvard** Steven A. McCarroll (Principal Investigator) [13,40], Robert E. Handsaker (Project Leader)[13,40], David M. Altshuler[3], Eric Banks[13], Guillermo del Angel[13], Giulio Genovese[13], Chris Hartl[13], Heng Li[13], Seva Kashin[13,40], James C. Nemesh[13], Khalid Shakir[13], **Cold Spring Harbor Laboratory** Seungtai C. Yoon (Principal Investigator)[42], Jayon Lihm[42], Vladimir Makarov[43], **Cornell University** Jeremiah Degenhardt[7], **European Molecular Biology Laboratory** Jan O. Korbel (Principal Investigator) (Co-Chair)[12,17], Markus H. Fritz[46], Sascha Meiers[17], Benjamin Raeder[17], Tobias Rausch[17,46], Adrian M. Stütz[17], **European Molecular Biology Laboratory, European Bioinformatics Institute** Paul Flicek (Principal Investigator)[12], Francesco Paolo Casale[12], Laura Clarke[12], Richard E. Smith[12], Oliver Stegle[12], Xiangqun Zheng-Bradley[12], **Illumina** David R. Bentley (Principal Investigator)[5], Bret Barnes[39], R. Keira Cheetham[5], Michael Eberle[5], Sean Humphray[5], Scott Kahn[39], Lisa Murray[5], Richard Shaw[5], **Leiden University Medical Center,** Eric-Wubbo Lameijer[118], **Louisiana State University** Mark A. Batzer (Principal Investigator)[52], Miriam K. Konkel[52], Jerilyn A. Walker[52], **McDonnell Genome Institute at Washington University** Li Ding (Principal Investigator)[22], Ira Hall[22], Kai Ye[22], **Stanford University** Phil Lacroute[65], **The Jackson Laboratory for Genomic Medicine** Charles Lee (Principal Investigator) (Co-Chair)[18,19], Eliza Cerveira[18], Ankit Malhotra[18], Jaeho Hwang[18], Dariusz Plewczynski[18], Kamen Radew[18], Mallory Romanovitch[18], Chengsheng Zhang[18], **Translational Genomics Research Institute** David W. Craig (Principal Investigator)[74], Nils Homer[75], **US National Institutes of Health** Deanna Church[34], Chunlin Xiao[25], **University of California, San Diego** Jonathan Sebat (Principal Investigator)[77], Danny Antaki[77], Vineet Bafna[119], Jacob Michaelson[120], Kenny Ye[79], **University of Maryland School of Medicine** Scott E. Devine (Principal Investigator)[90], Eugene J. Gardner (Project Leader)[90], **University of Michigan** Gonçalo R. Abecasis (Principal Investigator)[2], Jeffrey M. Kidd (Principal Investigator)[91,92], Ryan E. Mills (Principal Investigator)[91,92], Gargi Dayama[91,92], Sarah Emery[92], Goo Jun[2,94], **University of North Carolina at Charlotte** Xinghua Shi (Principal Investigator)[102], Andrew Quitadamo[102], **University of Oxford** Gerton Lunter (Principal Investigator)[8], Gil A. McVean (Principal Investigator)[8,9], **University of Texas MD Anderson Cancer Center** Ken Chen (Principle Investigator)[121], Xian Fan[121], Zechen Chong[121], Tenghui Chen[121], **University of Utah** David Witherspoon[105], Jinchuan Xing[106], **University of Washington** Evan E. Eichler (Principal Investigator) (Co-Chair)[10,11], Mark J. Chaisson[10], Fereydoun Hormozdiari[10], John Huddleston[10,11], Maika Malig[10], Bradley J. Nelson[10], Peter H. Sudmant[10], **Vanderbilt University School of Medicine** Nicholas F. Parrish[95], **Weill Cornell Medical College**, Ekta Khurana (Principal Investigator)[109], **Wellcome Trust Sanger Institute** Matthew E. Hurles (Principal Investigator)[4], Ben Blackburne[4], Sarah J. Lindsay[4], Zemin Ning[4], Klaudia Walter[4], Yujun Zhang[4], **Yale University** Mark B. Gerstein (Principal Investigator)[113-115], Alexej Abyzov[116], Jieming Chen[113], Declan Clarke[117], Hugo Lam[122], Xinmeng Jasmine Mu[13,113], Cristina Sisu[113], Jing Zhang[113,115], Yan Zhang[113,115]

Exome Group: Baylor College of Medicine Richard A. Gibbs (Principal Investigator) (Co-Chair)[14], Fuli Yu (Project Leader)[14], Matthew Bainbridge[14], Danny Challis[14], Uday S. Evani[14], Christie Kovar[14], James Lu[14], Donna Muzny[14], Uma Nagaswamy[14], Jeffrey G. Reid[14], Aniko Sabo[14], Jin Yu[14], **BGI-Shenzhen** Xiaosen Guo[26,27], Wangshen Li[26], Yingrui Li[26], Renhua Wu[26], **Boston College** Gabor T. Marth (Principal Investigator) (Co-Chair)[23], Erik P. Garrison[4], Wen Fung Leong[23], Alistair N. Ward[23], **Broad Institute of MIT and Harvard** Guillermo del

Angel[13], Mark A. DePristo[41], Stacey B. Gabriel[13], Namrata Gupta[13], Chris Hartl[13], Ryan E. Poplin[13], **Cornell University** Andrew G. Clark (Principal Investigator)[7], Juan L. Rodriguez-Flores[45], **European Molecular Biology Laboratory, European Bioinformatics Institute** Paul Flicek (Principal Investigator)[12], Laura Clarke[12], Richard E. Smith[12], Xiangqun Zheng-Bradley[12], **Massachusetts General Hospital** Daniel G. MacArthur (Principal Investigator)[53], **McDonnell Genome Institute at Washington University** Elaine R. Mardis (Principal Investigator)[22], Robert Fulton[22], Daniel C. Koboldt[22], **Stanford University** Carlos D. Bustamante (Principal Investigator)[65], Simon Gravel[54], **Translational Genomics Research Institute** David W. Craig (Principal Investigator)[74], Alexis Christoforides[74], Nils Homer[75], Tyler Izatt[74], **US National Institutes of Health** Stephen T. Sherry (Principal Investigator)[25], Chunlin Xiao[25], **University of Geneva** Emmanouil T. Dermitzakis (Principal Investigator)[87-89], **University of Michigan** Gonçalo R. Abecasis (Principal Investigator)[2], Hyun Min Kang[2], **University of Oxford** Gil A. McVean (Principal Investigator)[8,9], **Yale University** Mark B. Gerstein (Principal Investigator)[113-115], Suganthi Balasubramanian[115], Lukas Habegger[113]

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

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

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

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

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

(PEL) Carla Gallo[163], Giovanni Poletti[163], **Punjabi in Lahore, Pakistan (PJL)** Danish Saleheen[164,165], Asif Rasheed[164]

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

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

1 Department of Genetics, Albert Einstein College of Medicine, Bronx, New York 10461, USA.
2 Center for Statistical Genetics, Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA.
3 Vertex Pharmaceuticals, Boston, MA 02210, USA.
4 Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK.
5 Illumina United Kingdom, Chesterford Research Park, Little Chesterford, Nr Saffron Walden, Essex CB10 1XL, UK.
6 McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA.
7 Center for Comparative and Population Genomics, Cornell University, Ithaca, New York 14850, USA.
8 Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK.
9 Department of Statistics, University of Oxford, Oxford OX1 3TG, UK.
10 Dept of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA.
11 Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA.
12 European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.
13 The Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA.
14 Baylor College of Medicine, Human Genome Sequencing Center, Houston, Texas 77030, USA.
15 US National Institutes of Health, National Human Genome Research Institute, 31 Center Drive, Bethesda, Maryland 20892, USA.
16 Centre of Genomics and Policy, McGill University, Montreal, Quebec H3A 1A4, Canada.
17 European Molecular Biology Laboratory, Genome Biology Research Unit, Meyerhofstr. 1, Heidelberg, Germany.
18 The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, Connecticut 06032, USA.
19 Department of Life Sciences, Ewha Womans University, Ewhayeodae-gil, Seodaemun-gu, Seoul, South Korea 120-750.
20 Max Planck Institute for Molecular Genetics, D-14195 Berlin-Dahlem, Germany.
21 Dahlem Centre for Genome Research and Medical Systems Biology, D-14195 Berlin-Dahlem, Germany.

22 McDonnell Genome Institute at Washington University, Washington University School of Medicine, St Louis, Missouri 63108, USA.
23 USTAR Center for Genetic Discovery & Department of Human Genetics, University of Utah School of Medicine.
24 Affymetrix, Inc., Santa Clara, California 95051, USA.
25 US National Institutes of Health, National Center for Biotechnology Information, 45 Center Drive, Bethesda, Maryland 20892, USA.
26 BGI-Shenzhen, Shenzhen 518083, China.
27 Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen, Denmark.
28 Princess Al Jawhara Albrahim Center of Excellence in the Research of Hereditary Disorders, King Abdulaziz University, Jeddah, Saudi Arabia.
29 Macau University of Science and Technology, Avenida Wai long, Taipa, Macau 999078, China.
30 Department of Medicine and State Key Laboratory of Pharmaceutical Biotechnology, University of Hong Kong, 21 Sassoon Road, Hong Kong.
31 Coriell Institute for Medical Research, Camden, New Jersey 08103, USA.
32 European Centre for Public Heath Genomics, UNU-MERIT, Unsiversity Maastricht,PO Box 616, 6200 MD Maastricht, The Netherlands.
33 Alacris Theranostics GmbH, D-14195 Berlin-Dahlem, Germany.
34 Personalis, Inc., Menlo Park, California 94025, USA.
35 US National Institutes of Health, National Human Genome Research Institute, 50 South Drive, Bethesda, Maryland 20892, USA.
36 Dept of Computer Engineering, Bilkent University, TR-06800 Bilkent, Ankara, Turkey.
37 Seven Bridges Genomics, Inc., 1 Broadway, 14th Floor, Cambridge, MA 02142, USA.
38 University of Oklahoma.
39 Illumina, Inc., San Diego, California 92122, USA.
40 Dept of Genetics, Harvard Medical School, Cambridge, Massachusetts 02142, USA.
41 SynapDx, Four Hartwell Place, Lexington, MA 02421, USA.
42 Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA.
43 Seaver Autism Center and Dept of Psychiatry, Mount Sinai School of Medicine, New York, New York 10029, USA.
44 Dept of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853, USA.
45 Department of Genetic Medicine, Weill Cornell Medical College, New York, NY, 10044, USA.
46 European Molecular Biology Laboratory, Genomics Core Facility, Meyerhofstrasse 1, 69117 Heidelberg, Germany.
47 Bill Lyons Informatics Centre, UCL Cancer Institute, University College London, London WC1E 6DD, UK.
48 Center for Systems Biology and Dept Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA.
49 Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA.
50 Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK.
51 Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1498, New York, NY 10029-6574, USA.

52 Dept of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana 70803, USA.
53 Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA.
54 McGill University and Genome Quebec Innovation Centre, 740, Avenue du Dr. Penfield, Montreal, Qc, Canada.
55 National Eye Institute, National Institutes of Health, Bethesda, Maryland, 20892.
56 New York Genome Center, 101 Avenue of the Americas, 7th floor, New York, NY 10013, USA.
57 Department of Systems Biology, Columbia University, New York, NY 10032, USA.
58 Department of Computer Science, Fu Foundation School of Engineering, Columbia University, New York, NY, USA.
59 Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA USA.
60 General Hospital and Harvard Medical School, Boston, MA USA.
61 Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142, USA.
62 Ontario Institute for Cancer Research, MaRS Centre, 661 University Avenue, Suite 510, Toronto, Ontario, M5G 0A3, Canada.
63 Dept of Anthropology, Penn State University, University Park, Pennsylvania 16802, USA.
64 Rutgers Cancer Institute of New Jersey, New Brunswick, NJ 08903, USA.
65 Dept of Genetics, Stanford University, Stanford, California 94305, USA.
66 Departments of Genetics and Pathology, Stanford University, Stanford, California 94305-5324, USA.
67 Ancestry.com, San Francisco, California 94107, USA.
68 DNAnexus, 1975 W El Camino Real STE 101, Mountain View CA 94040, USA.
69 Laboratorio Nacional de Genómica para la Biodiversidad (LANGEBIO), CINVESTAV, Irapuato, Guanajuato 36821, Mexico.
70 Blavatnik School of Computer Science, Tel-Aviv University, Israel, 69978.
71 Dept of Microbiology, Tel-Aviv University, Israel, 69978.
72 International Computer Science Institute, Berkeley, California 94704, USA.
73 Thermo Fisher Scientific, 200 Oyster Point Boulevard, South San Francisco, CA 94080, USA.
74 The Translational Genomics Research Institute, Phoenix, Arizona 85004, USA.
75 Life Technologies, Beverly, Massachusetts 01915, USA.
76 Dept of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, California 90024, USA.
77 Dept of Psychiatry, University of California, San Diego, La Jolla, California 92093, USA.
78 Dept of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, California 92093, USA.
79 Dept of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, New York 10461, USA.
80 Depts of Bioengineering & Therapeutic Sciences, University of California, San Francisco, San Francisco, California 94158, USA.
81 Institute for Quantitative Biosciences (QB3), University of California, San Francisco, 1700 4th Street San Francisco, California 94158.
82 Institute for Human Genetics, University of California, San Francisco, 1700 4th Street San Francisco, California 94158.
83 Center for Biomolecular Science and Engineering, University of California-Santa Cruz, Santa Cruz, California 95064, USA.

84 Howard Hughes Medical Institute, Santa Cruz, California 95064, USA.
85 Dept of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA.
86 Dept of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK.
87 Dept of Genetic Medicine and Development, University of Geneva Medical School, Geneva, 1211 Switzerland.
88 Institute for Genetics and Genomics in Geneva, University of Geneva, 1211 Geneva, Switzerland.
89 Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland.
90 Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA.
91 Dept of Computational Medicine and Bioinfomatics, University of Michigan, Ann Arbor, Michigan 48109, USA.
92 Department of Human Genetics, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA.
93 Department of Pediatrics, University of Pittsburgh, Pittsburgh, PA 15224.
94 The University of Texas Health Science Center at Houston, Houston, Texas 77030, USA.
95 Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA.
96 University of Michigan Sequencing Core, University of Michigan, Ann Arbor, Michigan 48109, USA.
97 Istituto di Ricerca Genetica e Biomedica, CNR, Monserrato, 09042 Cagliari, Italy.
98 Dipartimento di Scienze Biomediche, Università delgi Studi di Sassari, 07100 Sassari, Italy.
99 UT Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390.
100 Dept of Pediatrics, University of Montreal, Ste. Justine Hospital Research Centre, Montreal, Quebec H3T 1C5, Canada.
101 Department of Genetics, Department of Biostatistics, Department of Computer Science, University of North Carolina, Chapel Hill 27599.
102 Department of Bioinformatics and Genomics, College of Computing and Informatics, University of North Carolina at Charlotte, 9201 University City Blvd., Charlotte, NC 28223, USA.
103 Department of Medical Genetics, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, The Netherlands.
104 Dept of Biology, University of Puerto Rico at Mayagüez, Mayagüez, Puerto Rico 00680, USA.
105 Eccles Institute of Human Genetics, University of Utah School of Medicine, Salt Lake City, Utah 84112, USA.
106 Dept of Genetics, Rutgers University, Piscataway, New Jersey 08854, USA.
107 Dept of Medicine, Division of Medical Genetics, University of Washington, Seattle, Washington 98195, USA.
108 Department of Biostatistics, University of Washington, Seattle, Washington 98195, USA.
109 Department of Physiology and Biophysics, Weill Cornell Medical College, New York, NY, 10065.
110 Department of Human Genetics, Radboud Institute for Molecular Life Sciences and Donders Centre for Neuroscience, Radboud University Medical Center, Geert Grooteplein 10, 6525 GA Nijmegen.
111 Department of Molecular Developmental Biology, Faculty of Science, Radboud Institute for Molecular Life Sciences (RIMLS), Radboud University, 6500 HB Nijmegen, The Netherlands.

112 Institute of Genetics and Biophysics, National Research Council (CNR), 80125 Naples, Italy.
113 Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA.
114 Dept of Computer Science, Yale University, New Haven, Connecticut 06520, USA.
115 Dept of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA.
116 Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905.
117 Dept of Chemistry, Yale University, New Haven, Connecticut 06520, USA.
118 Molecular Epidemiology Section, Dept of Medical Statistics and Bioinformatics, Leiden University Medical Center 2333 ZA, The Netherlands.
119 Dept of Computer Science, University of California, San Diego, La Jolla, California 92093, USA.
120 Beyster Center for Genomics of Psychiatric Diseases, University of California-San Diego, La Jolla, California 92093, USA.
121 Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77230, USA.
122 Bina Technologies, Roche Sequencing, Redwood City, CA, 94065, USA.
123 Department of Surgery, Massachusetts General Hospital, Boston, Massachusetts 02114, USA.
124 Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kemitorvet Building 208, 2800 Lyngby, Denmark.
125 Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY.
126 Department of Invertebrate Zoology, American Museum of Natural History, New York, New York 10024, USA.
127 School of Life Sciences, Arizona State University, Tempe, AZ 85287-4701, USA.
128 Program in Biomedical Informatics, Stanford University, Stanford, CA 94305, USA.
129 Institute for Molecular Bioscience, University of Queensland, St Lucia, QLD 4072, Australia.
130 Virginia Bioinformatics Institute, 1015 Life Sciences Drive, Blacksburg, VA 24061, USA.
131 Division of Allergy & Clinical Immunology, School of Medicine, Johns Hopkins University, Baltimore, Maryland 21205, USA.
132 Dept of Ecology and Evolution, Stony Brook University, Stony Brook NY 11794.
133 Centre for Health, Law and Emerging Technologies, University of Oxford, Oxford OX3 7LF, UK.
134 Genetic Alliance, London, N1 3QP, UK.
135 The Ethox Center, Nuffield Department of Population Health, University of Oxford, Old Road Campus, OX3 7LF.
136 Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA.
137 Dept of Medical History and Bioethics, Morgridge Institute for Research, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA.
138 University of Wisconsin Law School, Madison, Wisconsin 53706, USA.
139 US National Institutes of Health, Center for Research on Genomics and Global Health, National Human Genome Research Institute, 12 South Drive, Bethesda, Maryland 20892, USA.
140 Department of African & African American Studies, Duke University, Durham, North Carolina 27708, USA.
141 Dept of Genetics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA.

142 Department of Psychiatry and Clinical Psychobiology & Institute for Brain, Cognition and Behavior (IR3C), University of Barcelona, 08035 Barcelona, Spain.
143 Cancer and Immunogenetics Laboratory, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DS, UK.
144 Laboratory of Molecular Genetics, Institute of Biology, University of Antioquia, Medellín, Colombia.
145 Peking University Shenzhen Hospital, Shenzhen, 518036, China.
146 Institute of Medical Biology, Chinese Academy of Medical Sciences & Peking Union Medical College, Kunming 650118, China.
147 Instituto de Biologia Molecular y Celular del Cancer, Centro de Investigacion del Cancer/IBMCC (CSIC-USAL), Institute of Biomedical Research of Salamanca (IBSAL) & National DNA Bank Carlos III, University of Salamanca, Salamanca, Spain.
148 Ponce Research Institute, Ponce Health Sciences University, Ponce, Puerto Rico, 00716.
149 Chronic Disease Research Centre, Tropical Medicine Research Institute, Cave Hill Campus, The University of the West Indies.
150 Faculty of Medical Sciences, Cave Hill Campus, The University of the West Indies.
151 Tropical Metabolism Research Unit, Tropical Medicine Research Institute, Mona Campus, The University of the West Indies.
152 International Centre for Diarrhoeal Disease Research, Dhaka, Bangladesh.
153 Xishuangbanna Health School, Xishuangbanna 666100, China.
154 Irrua Specialist Teaching Hospital, Edo State, Nigeria.
155 Redeemers University, Ogun State, Nigeria.
156 Harvard T.H. Chan School of Public Health, Boston, Massachusetts 02115, USA.
157 Medical Research Council Unit, The Gambia.
158 NHLI, Imperial College London, Hammersmith Hospital, London, United Kingdom.
159 Centre for Tropical Medicine, Oxford University Clinical Research Unit, Ho Chi Minh City, Viet Nam.
160 Peter Doherty Institute of Infection and Immunity, The University of Melbourne, Australia.
161 Kenema Government Hospital, Ministry of Health and Sanitation, Kenema, Sierra Leone.
162 Tulane University Health Sciences Center, New Orleans, USA.
163 Laboratorios de Investigación y Desarrollo, Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Peru.
164 Center for Non-Communicable Diseases, Karachi, Pakistan.
165 Dept of Epidemiology and Biostatistics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104.
166 US National Institutes of Health, National Human Genome Research Institute, 5635 Fishers Lane, Bethesda, Maryland 20892, USA.
167 Wellcome Trust, Gibbs Building, 215 Euston Road, London NW1 2BE, UK.
168 James D. Watson Institute of Genome Sciences, Hangzhou 310008, China.