doi:10.1038/nature09708

Supplementary Methods

SV detection with diverse approaches

We applied nineteen different SV discovery algorithms (*i.e.*, methods) – i.e. 6 RP, 4 RD, 3 SR, 4 AS, and 2 PD methods – to DNA sequence data from 185 individuals from the 1000GP pilot phase. The SV discovery methods were, in part, optimized towards discovering SVs in data from particular sequencing technologies (Illumina Solexa, Roche 454, and Life Technologies SOLiD), and for detecting SVs in sequence data with varying sequencing coverage (2-30X) and physical coverage (i.e., coverage in terms of spanning read-pairs; see Supplementary Table 1). For example, two RP methods operated with paired-ends generated with long insert size (greater than 1 kb insert size). In the figures, these RP methods are referred to as 'RL' methods, for "RP long" (e.g., Fig. 2A suggests marked SV ascertainment differences between RP and RL). We unified and harmonized the reported coordinates of candidate SV events from each individual discovery method by correcting them according to a leftmost breakpoint convention and converting the methods' callsets into a standardized, custom format for downstream validation and analysis. We filtered out any discovered deletions smaller than 50bp (reported elsewhere[33]) based on a first pass analysis, and did not re-assess those candidates following breakpoint assembly. The specific methods and parameters used in this project are listed in Supplementary Table 1.

Systematic validation with PCR and microarrays to assign FDRs using a hierarchical approach

PCR primers were designed for randomly (without replacement) chosen SV predictions from each callset using an automated primer design pipeline. The primer design pipeline involved the primer3 algorithm (available from http://frodo.wi.mit.edu/primer3/) for primer placement, and in-silico PCR (available from http://genome.ucsc.edu/cgi-bin/hgPcr) applied with default parameters to confirm proper placements. Primer pairs generating unique amplicons were kept and used in the PCR experiments. If primer pairs generated more than one amplicon at the given size (or at a smaller size) the primer positions were masked with 'N's and the primer design pipeline was re-initiated on the masked sequence. If primer3 failed to identify suitable primers, the windows for primer design were iteratively increased from 150bp flanking either breakpoint confidence interval end in steps of 150bp up to a maximum of 2kb. PCR experiments were carried out using previously described protocols[33]. Successfully amplified PCR products were assessed for the presence or absence of the expected alternative allele, either visually, or through chain termination sequencing of PCR products with a capillary sequencer.

Custom array-CGH DNA Microarrays[1] were used to validate deletions and duplications in the high coverage trios. The high resolution probe design allowed for the direct interrogation of probes falling into predicted SV candidate regions. Models of probe intensity in regions with expected copy number of 2 and non-2 were built from regions previously assessed[13]. Two different methods utilized these models to interrogate each region. The first used an 'alternative models' approach whereby a log odds score was calculated for each copy-number state, and the most probable state was inferred based on a given threshold. This approach had the ability to either validate or invalidate a region. The

second approach used a 'deviation from the null' approach, whereby an empirical distribution was constructed from the intensity model of regions with known copy number 2, and probes at both tails were interrogated in such a way that regions significantly deviating from this distribution were considered as validated[33]. We note that due to the comparative nature of CGH, validation with array-CGH may overestimate FDR, as real SVs in studied samples can be invalidated if they are also present in the array reference individual, thus resulting in the same/similar probe intensities in the SV region[45].

In addition to the custom CGH arrays applied to the high coverage trios, a combination of standard array-CGH and SNP microarrays were used to validate variants in the individuals sequenced at low coverage. A virtual 'superarray' was constructed from 156 Affymetrix 6.0 and Illumina 1M arrays used as part of the HapMap Project (http://hapmap.ncbi.nlm.nih.gov/), as well as standard Nimblegen 2.1M arrays. We developed a non-parametric test based on the simple assumption that, for any probe, samples with lower underlying copy number will, on average, tend to have a lower intensity measurement than samples with greater underlying copy number. The Wilcox rank sum test was applied to probe intensities for each probe falling inside putative SV regions. Samples were ranked in intensity space, then the ranks of all probes for samples with inferred copy-number of '2' samples were compared to the ranks of all other samples. Rank data across all the probes within the putative deletion or duplication were then combined. Putative deletions and duplications were considered validated if a significant Wilcox rank sum P-value of $P < 0.01$ was measured.

An assembly-based approach was also used to validate deletion calls through reconstruction of the breakpoint sequences. Illumina read-pairs overlapping intervals of (-500bp, +50bp) and (-100bp, +450bp) around the predicted start and end positions of each variant, respectively, were obtained from the sequence alignment (BAM format) files of each respective individual using samtools[46]. Each set of reads was analyzed and assembled using the TIGRA targeted assembly method. We applied cross-match to align each of the resulting contigs to the reference sequence at the associated variant location +/- 500bp flanking regions, yielding the SV breakpoints. A call was considered validated if an SV consistent with breakpoint confidence intervals of the candidate SV was observed in the alignment results.

FDRs were initially calculated separately for each individual experimental validation approach (i.e., PCR and array-CGH). FDRs were determined by dividing the number of invalidated calls by the total number of calls assessable by a particular validation platform. The estimated FDR for callsets validated by the superarray method was estimated as two times the fraction of putative calls for which we measured a Wilcox rank sum P-value of $P > 0.5$. FDRs estimates obtained from PCR and array-CGH were in good agreement (Supplementary Tables 2 and 3).

We implemented a hierarchal approach to combine these PCR and array-based FDRs as follows. We first used the validation platform for which the largest fraction of interpretable results were observed and weighted its FDR by this amount; the FDR of the latter was weighted according the remaining fraction of all events. These FDRs were then summed together to construct the final callset FDR. Thus, for *N* total calls in the trio data,

$$FDR = \frac{CGH_{invalidated}}{CGH_{validated} + CGH_{invalidated}} * \frac{CGH_{validated} + CGH_{invalidated}}{N} + \frac{PCR_{invalidated}}{PCR_{validated} + PCR_{invalidated}} * (1 - \frac{(CGH_{validated} + CGH_{invalidated})}{N})$$

A similar approach was used in the low-coverage data, with the substitution of CGH FDR with FDR from the superarray validation.

As we were estimating the FDR of SV discovery methods we also recorded the rate at which SVs falling into different size spectra were validated (Supplementary Figure 11), and observed an overall uniform validation rate across the SV size spectrum considered in our survey.

Validation by microarray-based sequence capture

We attempted to also validate 2,414 regions, for which deletions were predicted in NA12878, using a microarray-based sequence capture approach. For this purpose a custom Nimblegen microarray with probes covering 2 kb flanking regions of deletion breakpoints was designed. Array design was optimized to maximize the uniform coverage over target regions by using probes of ~75 bp in length containing unambiguously mappable sequence (i.e. the probe sequences have a single hit in build 36 of the human reference genome). Overall, 65-82% of target regions were covered by probes. Genomic DNA from three samples corresponding to a patent offspring trio with European ancestry (daughter NA12878, mother NA12892, and father NA12891) was hybridized to the array. Captured DNA was sequenced using the 454 GS FLX Titanium platform, yielding approximately ~1x coverage per haplotype per sample.

Reads were aligned to the human reference genome using Megablast and those mapped to the target regions were subsequently realigned using the Needleman-Wunsch algorithm with zero gap extension penalty (in order to allow for alignment extension across large gaps). Needleman-Wunsch alignments were post-processed by merging alignment fragments separated by less than 5 bp gaps and by removing fragments shorter than 20 bp. The breakpoints flanking the largest gap were compared to the predicted deletion breakpoints to validate the deletion.

Estimation of method sensitivity using gold standards

Gold standard data sets were constructed for NA12878 and NA12156 from both published data[1,13,22], and unpublished SVs mapped at nucleotide resolution using previously described approaches[22,28]. Sensitivity was calculated for each callset by dividing the number of overlapping events (1bp overlap criteria) by the total number of calls in the gold standard set.

Obtaining breakpoint residuals and precision-aware merging of SVs

The breakpoint resolution of each callset was determined by comparing deletion calls with assembled breakpoints derived from the assembly-based validation with TIGRA. We required a 50% reciprocal overlap between deletion calls and the respective assembled deletions as a pre-filter to avoid comparing calls that correspond to different deletions. Start and end position residuals were obtained from each matched call, resulting in the distributions of deviations from the actual event (Supplementary Fig. 2). These residuals were used to estimate confidence intervals for each deletion call. The confidence intervals were characterized by the high and low extent from the mode (CI+ for the extent to the right and CI- to the left). The most probable position was estimated as:

$$startbreak = Start_{call}\text{-}Start_{offset}$$

$$endbreak = End_{call}\text{-}End_{offset}$$

While the confidence intervals were bracketed by:

$$(Start_{call} - Start_{CI+}) < Start < (Start_{call} + Start_{CI-})$$
$$(End_{call} - End_{CI+}) < End < (End_{call} + End_{CI-})$$

In some cases the confidence intervals were significantly larger than the size of the called deletion. In these cases the confidence intervals were trimmed such that start CI could not extend beyond *endbreak* and the end CI was trimmed such that it would not precede *startbreak*.

The assessment of breakpoint accuracy for each call set facilitated the development of a novel, precision-based merging approach (Supplementary Fig. 5). We utilized this approach to merge calls between different callsets on the basis that an SV event independently discovered by two different algorithms should have a breakpoint falling within each of their intersecting confidence intervals. Thus, we required that SVs were merged together only if intersecting confidence intervals were displayed around each breakpoint. Then, the breakpoints were assigned using (*i*) assembled breakpoints, if available for one of the member calls or (*ii*) the midpoint of the intersecting intervals, with the upper and lower bounds of this intersection becoming the new interval for the merged call.

Detection methods for Tandem Duplications

Tandem duplication events represent a special case that can be observed as duplicated sequence in the reference absent in a sample ("t.dup deletion"), or as duplicated sequence in a sample where the reference genome has only one copy ("t.dup insertion"). Tandem duplication insertions were detected as clusters of read pair fragments spanning the breakpoint junction between the two duplicated regions, which appear to 'map backwards' (Fig 1a, RP green "Dup") in hg18 reference coordinates. Candidate events within known VNTR regions (RepeatMasker 3.27) were filtered out, along with events without evidence of increased depth of read coverage in the boundary of the duplication. Tandem duplication deletions were selected from the set of all 22025 deletions as events with a bracketing homologous region of approximately the same length as the deletion (within +/- 20bp).

Local assembly coordinates from TIGRA were used to identify the deletion breakpoints and to measure the homology length.

Assessment of novelty

The novelty of our release set was assessed by comparing the discovered variants to SVs reported in dbVAR (http://www.ncbi.nlm.nih.gov/dbvar; downloaded 6 June 2010), the Database of Genomic Variants[11] (DGV) (http://projects.tcag.ca/variation/; November 2010 set), as well as in various data sets from the analysis of individual genomic sequences[17,18,20,27,36,47,48,49,50,51] (some of the individual-specific datasets were available at DGV or dbVAR as well, but were added nonetheless as the databases did not represent these datasets across the entire SV size range ascertained by our study). We applied a 50% reciprocal overlap to determine whether a variant overlapped a previously reported SV, in order to infer novelty (Supplement).

We also investigated the stratification of SV calls classified as novel by both SV size and variant allele frequency for our genotyped deletions. As may be expected, both smaller sized as well as lower frequency SVs displayed the highest degree of novelty. Little novelty was identified amongst common SVs as well as amongst large SVs (Supplemental Fig. 15), consistent with a previous identification of these in earlier surveys.

Genotyping of deletions in low coverage sequences.

The Genome STRiP method, described in further detail elsewhere, was used to genotype discovered deletions in low coverage sequence data. In brief, this method utilized RD, SR, and RP features to assess a region for the presence or absence of a deletion. Likelihoods from RD, SR, and RP features were combined in a Bayesian model to generate initial genotype likelihoods. These initial likelihoods were then integrated with SNP haplotype information using Beagle (v3.1) and a reference panel of SNP genotypes from Hapmap3r2, to yield posterior genotype likelihoods for each deletion. The genotype refinement step was performed separately in each population; trio parents and children were analyzed separately. After the incorporation of haplotype information into posterior genotype likelihoods using Beagle, sites with sufficient information for genotyping were selected using two filters: (*i*) minimum call rate of 50% across all three populations using a genotype quality threshold of 13 (95% confidence) and (*ii*) Hardy-Weinberg equilibrium p-value > 0.01 in each of the three populations. Of note, these deletions where genotyped using a bi-allelic variant model that may generate rare inaccurate genotype calls in loci harboring both deletions and duplications (Supplement).

Formation mechanism analysis

SVs with breakpoints mapped at nucleotide resolution were analyzed with the BreakSeq classification pipeline and SVs were classified according to their likely mechanism of formation[41]. Furthermore, the ancestral states of the SVs were inferred by aligning breakpoint junction sequences to primate genomes as previously described[41].

A set of 22,373 recombination hotspots[52] were overlapped with the breakpoints of SVs formed by NAHR, NH, MEI, and VNTR. The probability of a random nucleotide overlapping

with a recombination hotspot region (a total of ~0.24 billion base pairs) in the human genome (~3 billion base pairs) was estimated to be 0.08. The *P*-value for the association of breakpoints with recombination hotspots for each SV mechanism was calculated assuming a binomial distribution.

We identified putative SV formation hotspots in the genome using a two-step process. As a first step, we disregarded BreakSeq SV formation mechanism classification by segmenting all deletions mapped to nucleotide resolution into potential SV hotspots. Partially overlapping deletions were made 'non-redundant' by removing the larger of two overlapping SVs for this analysis. Segmentation was performed by applying an SV coordinate-aware 500kb windowing approach and assessing the population of all *N* possible genomic windows containing SVs under a Poisson assumption (with *N* being the number of deletions assessed). Raw *P*-values were corrected using the Benjamini and Hochberg (BH) multiple testing correction, and more than fifty SV hotspots were predicted with $P_{BH}<0.01$. Adjacent windows with significant enrichment ($P_{BH}<0.01$) were joined. The predicted hotspots displayed enrichments in SV content over randomly shuffled SVs of five-fold or higher. As a second step, we classified SVs in each predicted SV hotspot according to their formation mechanism using BreakSeq. Most predicted hotspots were populated mainly by one SV formation mechanism (Fig. 5C).

Additional References of the Supplementary Methods

45 Park, H. *et al.* Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nat Genet* **42**, 400-405 (2010).

46 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

47 Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872-876, (2008).

48 Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60-65 (2008).

49 Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59 (2008).

50 Kim, J. I. *et al.* A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**, 1011-1015, (2009).

51 Lee, S., Hormozdiari, F., Alkan, C. & Brudno, M. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat Methods* **6**, 473-474 (2009).

52 Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**, 321-324 (2005).