Section 1: Supplementary Notes for "Mapping copy number variation by population scale genome sequencing"

Ryan E. Mills[1], Klaudia Walter[2], Chip Stewart[3], Robert E. Handsaker[4], Ken Chen[5], Can Alkan[6,7], Alexej Abyzov[8], Seungtai Chris Yoon[9], Kai Ye[10], R. Keira Cheetham[11], Asif Chinwalla[5], Donald F. Conrad[2], Yutao Fu[12], Fabian Grubert[13], Iman Hajirasouliha[14], Fereydoun Hormozdiari[14], Lilia M. Iakoucheva[15], Zamin Iqbal[16], Shuli Kang[15], Jeffrey M. Kidd[6], Miriam K. Konkel[17], Joshua Korn[4], Ekta Khurana[8,18], Deniz Kural[3], Hugo Y. K. Lam[13], Jing Leng[8], Ruiqiang Li[19], Yingrui Li[19], Chang-Yun Lin[20], Ruibang Luo[19], Xinmeng Jasmine Mu[8], James Nemesh[4], Heather E. Peckham[12], Tobias Rausch[21], Aylwyn Scally[2], Xinghua Shi[1], Michael P. Stromberg[3], Adrian M. Stütz[21], Alexander Eckehart Urban[13], Jerilyn A. Walker[17], Jiantao Wu[3], Yujun Zhang[2], Zhengdong D. Zhang[8], Mark A. Batzer[17], Li Ding[5,22], Gabor T. Marth[3], Gil McVean[23], Jonathan Sebat[15], Michael Snyder[13], Jun Wang[19,24], Kenny Ye[20], Evan E. Eichler[6,7], Mark B. Gerstein[8,18,25], Matthew E. Hurles[2], Charles Lee[1], Steven A. McCarroll[4,26], Jan O. Korbel[21], 1000 Genomes Project[#]

1. Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA
2. The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA UK.
3. Department of Biology, Boston College, Boston, MA
4. Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA
5. The Genome Center at Washington University, St. Louis, MO
6. Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA
7. Howard Hughes Medical Institute, University of Washington, Seattle, Washington, USA.
8. Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT
9. Seaver Autism Center and Department of Psychiatry, Mount Sinai School of Medicine, New York, NY
10. Departments of Molecular Epidemiology, Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, the Netherlands
11. Illumina Cambridge Ltd, Chesterford Research Park, Little Chesterford, Essex CB10 1XL, UK
12. Life Technologies, Beverly, MA
13. Department of Genetics, Stanford University, Stanford, CA
14. School of Computing Science, Simon Fraser University, Burnaby, British Columbia, Canada.
15. Department of Psychiatry, Department of Cellular and Molecular Medicine, Institute for Genomic Medicine, University of California, San Diego, La Jolla, CA
16. Wellcome Trust Centre for Human Genetics, University of Oxford, OX3 7BN, UK
17. Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana
18. Molecular Biophysics and Biochemistry Department, Yale University, New Haven, CT
19. BGI-Shenzhen, Shenzhen 518083, China
20. Albert Einstein College of Medicine, Bronx, NY
21. Genome Biology Research Unit, European Molecular Biology Laboratory, Heidelberg, Germany
22. Department of Genetics, Washington University, St. Louis, MO
23. Department of Statistics, University of Oxford, OX3 7BN, UK
24. Department of Biology, University of Copenhagen, Copenhagen, Denmark
25. Department of Computer Science, Yale University, New Haven, CT
26. Department of Genetics, Harvard Medical School, Boston, MA
[#]Lists of participants and affiliations appear in Supplementary Information.

November 25th, 2010

Contents

Abbreviations

Institutional

AB	Applied Biosystems / Life Technologies
AE	Albert Einstein College of Medicine
BC	Boston College
BG	Beijing Genomics Institute
BI	Broad Institute
LN	Leiden University Medical Center
OX	University of Oxford
SD	University of California, San Diego
SI	Wellcome Trust Sanger Institute
UW	University of Washington
WU	Washington University, St. Louis
YL	Yale University

Computational Approach

RP	Read Pair Mapping
RD	Read Depth Analysis
SR	Split Read Alignments
AS	Assembly of Sequence Reads
PD	Individual approaches integrating Read Pair and Read Depth features

Types of Structural Variation

DEL	Deletion
DUP	Duplication
INS	Insertion

List of Supplementary Figures

List of Supplementary Tables

Supplementary Notes

Algorithm-Centric Approach

Our analysis group's initial efforts focused on SV discovery with individual methods accompanied with extensive validations ("Release set"; Fig 1B in main text). We realized, however, that extensive validations are unlikely to be pursued regularly in future surveys of SVs in population scale sequence data. Thus, we also assessed an alternative approach for generating an SV discovery set; that is the "algorithm-centric set" involving only sparse validations (i.e., experiments at a scale sufficient for estimating the FDRs of the underlying SV discovery methods). The algorithm-centric set consisted of two steps. First, we collated SV calls generated by individual (single) methods achieving an FDR of 10% or less. Second, we systematically assessed whether and to what extent integrating results from complementary SV discovery methods (i.e., such utilizing distinct features, e.g. SR and RD) would allow for increased specificity and sensitivity in SV calling. To this end, we integrated callsets generated by individual (single) SV discovery methods in a pairwise fashion, i.e. based on pairwise intersection of the methods' callsets with our precision-aware merging approach (Supplementary Fig. 5). We reassessed the specificity of the resulting (pairwise intersected) callsets, and requested an FDR of 10% or less for inclusion of the resulting callsets in the algorithm-centric set. FDR re-evaluation was performed separately for each pairwise intersection of callsets using the hierarchical FDR estimation rationale described above.

We first assessed the accuracy of predicted deletions with supporting evidence from two methods with our pairwise callset intersection approach. Indeed, such deletion calls tended to have a higher accuracy (lower FDR) than calls generated with a single method and led us to consider predicted SVs from an additional sixteen individual SV discovery callsets that were not previously considered owing to their insufficient accuracy (compare Supplementary Table 15 with Supplementary Tables 3 and 4). The highest relative increases in accuracy were achieved when different SV discovery approaches, such as RD and RP approaches, were integrated in this fashion. Of note, the accuracy improved also when combining conceptually similar methods, such as when combining callsets from multiple RD-based methods. Furthermore, we observed an increase in accuracy compared to individual methods also when integrating two insertion callsets operating at a similar SV size range, i.e. callsets with predicted MEIs inferred with SR and RP methods, respectively.

The "algorithm-centric set" captured approximately two thirds of the discovery set generated by the "Release set" (compare Table 1 with Supplementary Table 7). Our deletion calls, generated with the "algorithm-centric set", further achieved a sensitivity of ~60% in the low-coverage data, and more than 80% in the trio data (Supplementary Fig. 4). We note that, unsurprisingly, algorithm-based SV discovery is particularly powerful in the trios, given the density of data in terms of coverage and the availability of data from diverse sequencing platforms, both of which the low-coverage genomes comprised to a lesser extent. We also assessed what fraction of the release set can be reconstructed when running optimal combinations of several algorithms on the sequence data, rather than all algorithms (Supplementary Tables 16, 17). In general, a large portion of our final callset (54%) would be missed if only the top two or three algorithms would be run by users. On the other hand, 48% of the release set can be reconstructed with the top five algorithms. We conclude that a considerable portion of our discovery set, but not its entirety, can be captured by combining computational methods with sparse validation (Supplementary Fig. 8).

Analysis of Callset Novelty

We also attempted to estimate to what extent callset novelty would be reduced through the introduction of data from individual genomes analyzed with short DNA read technologies. To this end we assessed callset novelty with a subset of previously reported SVs (i.e., such reported in dbVAR as well as in two individual genomes analyzed by long-read based sequence technology[1,2]), followed by the addition and assessment of data from additional personal genomes. For example, when specifically considering deletions reported in Bentley *et al.*[3] (who analyzed an anonymous African individual) we observed a 2% decrease in novel deletions; for McKernan *et al.*[4] (who analyzed the same individual as Bentley *et al.*) we also observed 2% decrease; for Kim *et al.*[5] (who analyzed an anonymous Korean individual) we observed a 0.003 decrease, and for Wang *et al.*[6] (who analyzed an anonymous East Asian individual) we observed a 0.2% decrease. We conclude – despite differences in variant callers and sequencing libraries/platforms amongst papers – that each additional genome decreases the detected novelty slightly, *i.e.*, by 0.2-2%, as rare SVs identified in our study are recapitulated in other genomes.

Breakdown of Microhomology at SV Breakpoints

10,125 of the 22,025 merged deletion calls – where a deletion was defined as a sequence loss relative to the human reference genome (build36/hg18) – had breakpoint assembly data at base-pair resolution. We initially analyzed the sequence context of these SVs by taking the 300bp flanking sequence of each deletion and searching for sequencing homology/microhomology. Altogether, we observed 7,717 SVs (76.2%) to be flanked by microhomology or homology stretches of 2bp to 376bp in length; 7.7% were blunt deletion; the remaining portion (16.1%) were deletions with recorded non-template sequence insertions. We stratified the microhomology analyses by ancestral status, and observed that the distribution of microhomology sequences peaks at 2bp for deletions relative to ancestral state (median=2bp), and at 15bp for insertions (median=9bp), as displayed in Supplementary Figure 9. Whereas the former are plausibly formed by NH, the latter were mostly associated with MEIs (i.e., *Alu* and LINE elements).

Analysis of Overlapping SVs with Distinct Breakpoints

We identified 497 deletions relative to the reference genome (4.4% of all deletions with breakpoint coordinates) for which different alternative breakpoint coordinates were inferred based on TIGRA targeted assembly of constituent callset-specific SVs (the whole list of candidate alternative breakpoints is in Supplementary Table 18). One potential explanation for this observation is the occurrence of distinct SV alleles at the same locus in the genome. We manually reviewed a number of these cases but didn't find an example with definitive evidence for multiple breakpoints. Thus, a more probably explanation is that many apparently overlapping SVs with different inferred breakpoints correspond to TIGRA mis-assemblies, as the rate of alternative SVs (4.4%) is within the estimated rate of TIGRA false-discoveries (*i.e.*, mis-assemblies) of 6.8% (assessed based on the validation of 151/162 TIGRA assemblies using PCR followed by Sanger sequencing). However, these loci did include twenty deletions intersecting with tandem duplications, i.e., loci that may represent multiallelic SVs or that may result from distinct interpretations as to whether a locus represents a deletion or duplication, based on method-specific reference sets applied (i.e., reference genome vs. population-reference).

Population Genetic Properties

The allelic state of each deletion in each genome was determined with the Genome STRiP method that combines information from RD, RP, SR and haplotype features of population scale sequence data to produce an allelic state determination (genotype) in each individual. This method was applied to discovered deletions in 156 individuals and was able to generate genotypes for 13,826 autosomal events. The genotypes displayed 99.1% concordance with array-based copy number genotypes (available for 1,970 of these deletions) from CGH arrays[7], indicating an excellent accuracy of the genotype data. Genome STRiP presumes a bi-allelic variant model, implying that it may generate rare inaccurate SV calls in multi-allelic loci harboring both deletions and duplications. Thus, a small number of individuals with copy number (CN) > 2 in multiallelic regions may be incorrectly genotyped as CN=2. Indeed, based on array-based analyses, the number of SVs that exhibit both loss and gain copy number appears to be modest: 5.4% of the autosomal genotyped CNVs in a recent array-CGH study[7] have a copy number less than two in some individuals and greater than two in others. Furthermore, only 161 out of 4899 (3.3%) of these same CNVs exhibit both gains and losses in the individuals analyzed by the 1000GP. Among these 161 variants – 55% of which our study rediscovered as deletions – we generally mis-called the individuals with >2 copies of a segment as CN=2. Note that based on array CGH data[7] the true average frequency of duplication copy numbers at these loci is 10.9% (thus, in the end, only very few genotypes are likely to be miscalled owing to the bi-allelic model, *i.e.* we expect that approximately 0.6% of our deletion calls may have been affected). Taken together, the abovementioned high concordance of Genome STRiP and the so far most detailed study of copy number genotypes[7] (an array-based study, which considered a multi-allelic variant model) suggests an overall low level of genotyping errors resulting from the bi-allelic variant model, while the bi-allelic model may explain some of the differences in concordance.

We assessed the taggability of genotyped deletions with SNPs, focusing on 5,068 deletions identified relative to the reference genome that displayed a minor allele frequency (MAF) >5%. Of these, 4,116 (81.2%) were tagged by at least one SNP. This figure is similar to the fraction of sequencing-derived SNPs that were reported to be well taggable by SNPs according to a recent study by the HapMap consortium[8]. Thus, it is likely that the poorly tagged

19% arise from factors that are not specific to SVs, such as the limited set of SNPs at each locus (that could serve as potential tags) and the uneven pattern of recombination across the genome (which gives rise to regions in which LD is relatively short in range). Nevertheless, we further examined factors that may influence taggability by relating deletion taggability to SV formation mechanism. This analysis revealed a slightly reduced LD for deletions formed by NAHR (only 77% of 190 common NAHR deletions had a tagSNP, *vs.* 81% of all deletions) and for deletions categorized as VNTR (66% of 145 common VNTR deletions had a tagSNP). We note that these observed slightly diminished taggability rates could in principle arise from recurrent SV formation, paucity of uniquely mapping SNPs to serve as potential tags, or modestly reduced genotyping accuracy at such sites. Furthermore, the LD properties of complex SVs (e.g., multiallelic SV), have not yet been fully ascertained as methods for genotyping such SVs with similar accuracy are still being developed.

Of note, the deletion genotype data further revealed nine apparently monomorphic deletions that appeared to be in the homozygous state in every individual interrogated. A closer examination suggested that four deletions spanned one or more contig gaps in the references assembly and thus were affected by non-reference bias (which led to the apparent monomorphic genotype). In addition, four deletions of smaller size (55bp – 5.6Kb) were in highly repetitive sequence; although the genotyped deletions are presumably present at high allele frequency, the better alignability of the non-reference allele may have confounded the determination of the actual allele frequency at these loci. The remaining deletion (chr18:5296843-5297330) appears to display a high allele frequency deletion involving a VNTR. While we did not observe any evidence for the reference allele in the low-coverage sequence data used for deletion genotyping, we did observe DNA reads supporting the reference allele in just one low coverage sample (NA12874) sequenced only with the 454 platform, suggesting that the reference genome contains a rare SV allele at this locus.

We also performed a principal component analysis (PCA) with the deletion genotype matrix, and display results showing the first two principal components of each genotyped sample in Supplementary Figure 10. As expected and previously reported based on array-CGH data[9], we observed a strong clustering of individuals within the same population and separation of each population-specific cluster. This implies an abundance of deletions in the genome that are based on relatively old SV formation events, with allele frequencies that have drifted independently in different populations resulting in the visible separation in the PCA. Additionally, we computed Wright's F_{ST} between each pair of populations (CEU, YRI and JPT+CHB) for each genotyped deletion to identify a possible enrichment of highly differentiated events in genic regions. Defining highly differentiated events as those with F_{ST} >= 0.5 in two of the three comparisons, we then performed a Chi^2 test on a 2x2 contingency table for different categories of genic overlap (including complete gene deletion and partial deletion intersection with coding sequence). We did not detect a statistically significant relationship between the high F_{ST} deletions and any category of genic overlap with this analysis (Supplementary Table 19).

Assessing *de novo* SV Formation in Parent-Offspring Trios

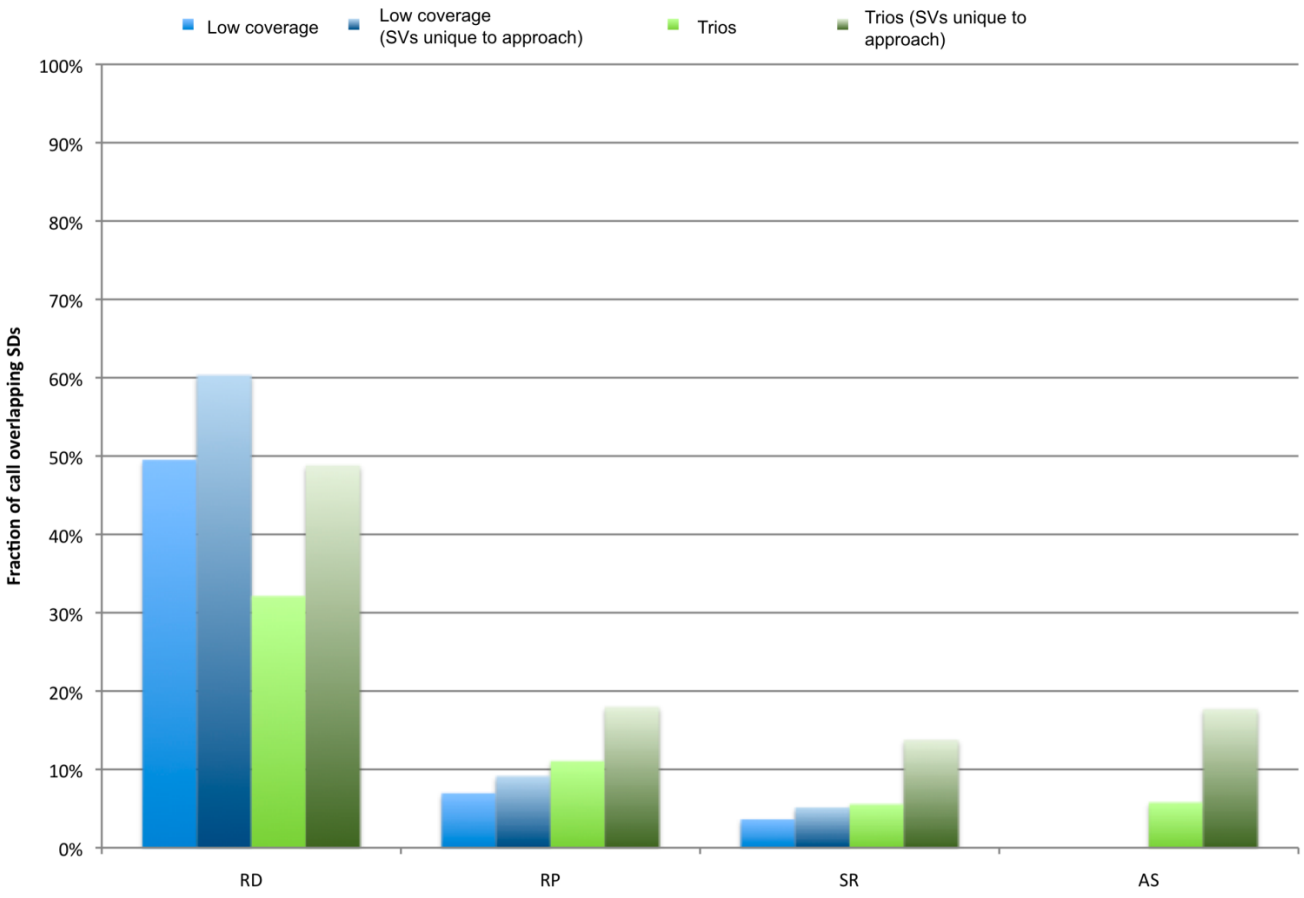
To identify putative *de novo* SV formation events we performed a detailed analysis of SVs in both of the parent-offspring trio. Specifically, we searched for patterns of disagreement in Mendelian inheritance and made use of 42M CGH arrays to gain additional support. We further used genotyping data from several sources to identify multi-allelic loci that may confound such analysis (more details on this analysis will be published elsewhere[10]). We applied stringent criteria for SV calling, as the detection of *de novo* SV formation events can be confounded by even a low FDR in SV discovery/genotyping. This analysis did not reveal any confident *de novo* SV, which may not be surprising given the recently estimated[7] low rate of *de novo* SV occurrence of $\mu = 3 \text{ X } 10^{-2}$.

Detection properties for Tandem Duplications

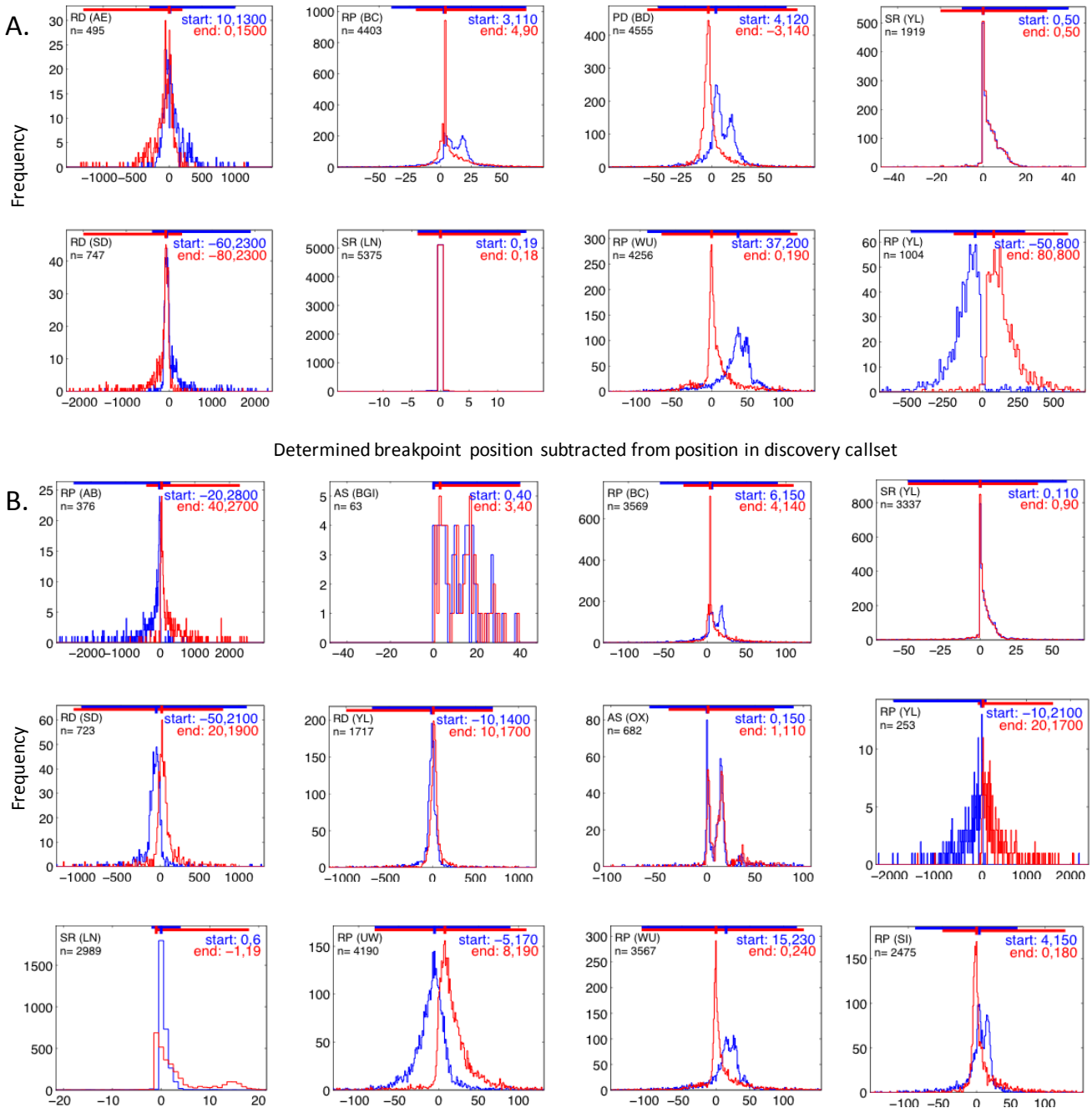
Tandem duplication events may be observed as back-to-back regions (length>50bp) of duplicated sequence in a sequenced sample where the reference genome has only one copy ("t.dup insertion") or as a deletion of one copy of a tandem duplicated region in the reference ("t.dup deletion"). We identified both types in the 1000 Genomes pilot data, and while both obviously refer to the same variant class mutation, the respective detection strategies were very different (with correspondingly different ascertainment biases). While false discovery rates for both types were less than 10% (Supplementary Table 2), it was difficult to access detection sensitivity due to the lack of available gold-standard tandem duplication events. The numbers of observed events (501 tandem duplications detected as *duplications* relative to the reference genome; 229 detected as *deletions* relative to the reference genome) can

therefore be interpreted as lower limits on the true numbers of tandem duplication events in the 1000 Genomes samples. We assessed the breakpoint resolution of the RP based tandem duplication detection algorithm (Supplementary Table 2) by comparing the predicted tandem duplication breakpoints with TIGRA targeted assembly as well as split-read based breakpoints (Supplementary Figure 14a). The tandem duplication length distribution ranges from 50 bp to 50kb (Supplementary Figure 14b). The abrupt limit of the length of tandem duplications detected as "deletions" compared to the reference genome arises from local assembly window around the candidate deletions, which as an approximation is roughly governed by the fragment length (~400bp) of the paired-end library insert size for most available sequencing reads,

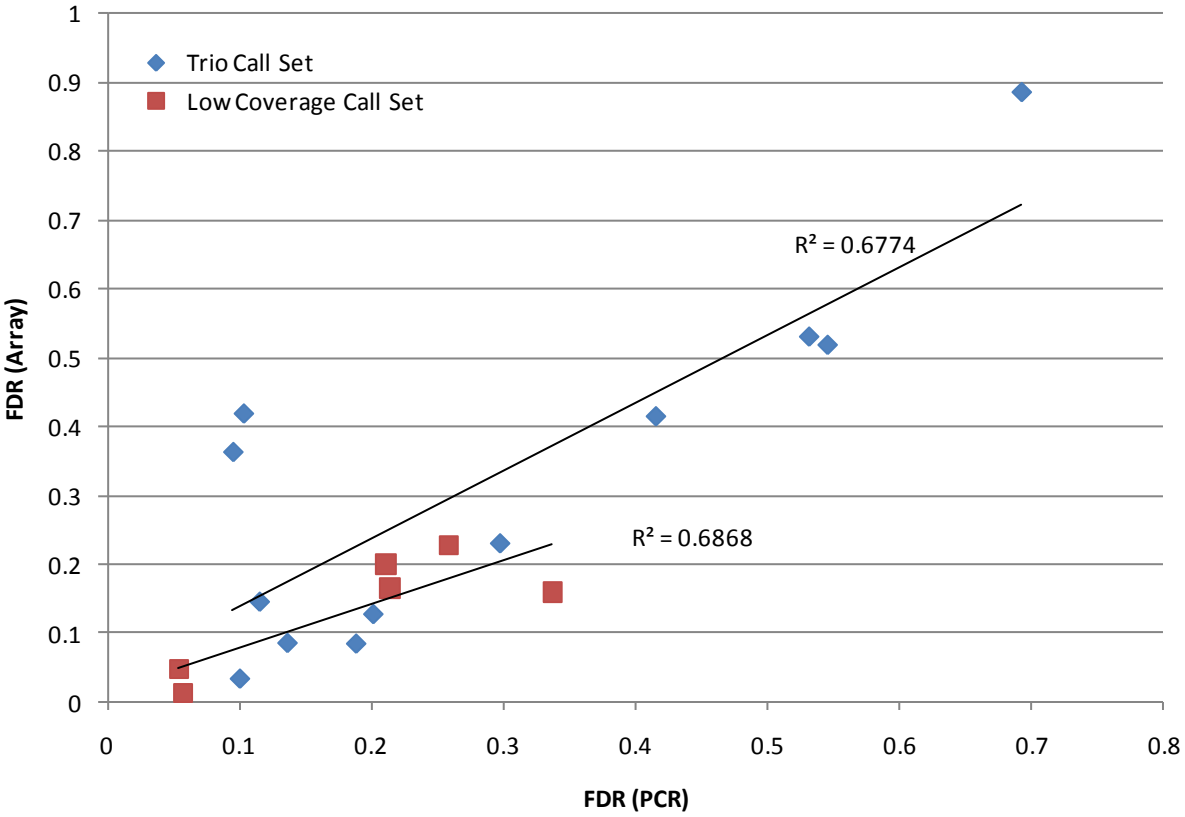
The ancestral allele for each tandem duplication was classified based on orthologous primate regions to determine if the human ancestor likely possessed the tandem duplication or was missing the duplication. We found that the ancestral allele was missing the duplication in more than 80% of the events ("ins" in Supplementary Figure 14c), but a significant number of events were unclear ("unc"). In several cases the chimpanzee, orangutan, & rhesus macaque genomes were split between the human reference allele and the alternate allele, indicating possible assembly biases in the primate genomes, such as the collapsing of tandem repeat units into a single sequence segment. Most tandem duplications exhibit microhomologies at the junction between the duplicated regions. Local assembly of the alternate allele sequence provided an estimate for the size of the microhomology. The tandem duplications displayed 2-17 bp of microhomology at their breakpoints (in ~80% of the cases), with the microhomology length 2 bp being most abundant. Furthermore, the tandem duplications did not display non-template inserted sequence.



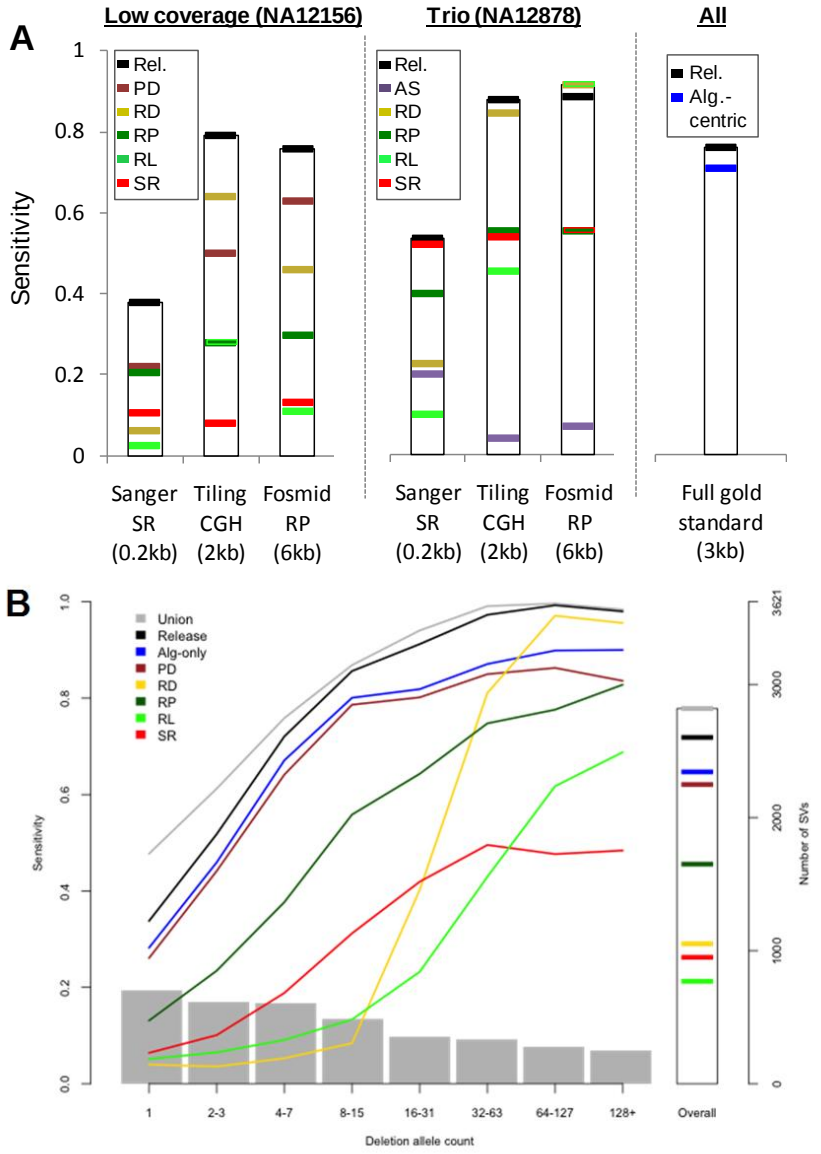
Supplementary Figure 1. Fraction of SV calls by each approach (RD, RP, SR, AS) intersecting segmental duplications (SD). SVs identified based on RD features (here displayed in terms of the union of RD methods) display a considerably stronger overlap with SDs than SVs identified by other methods. SD coordinates were downloaded from the UCSC browser. The extent of overlap was assessed by requiring an intersection of at least 1bp.



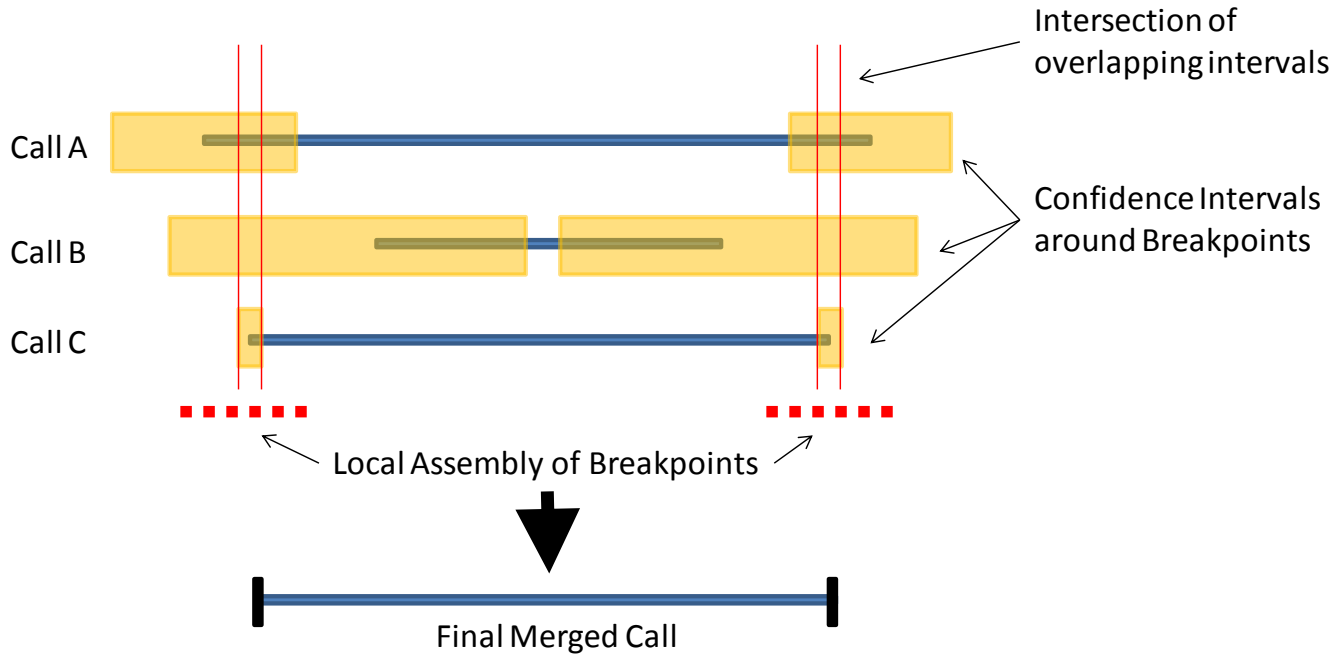
Supplementary Figure 2. Deletion callset breakpoint residual distributions for methods applied to (A) low coverage and (B) trio data. The blue and red histograms are the breakpoint residuals (callset-assembly) for the start and end coordinates of SVs called by the respective SV discovery methods. The horizontal lines at the top of each plot mark the 2% (2.3 sigma) confidence intervals. The vertical notches mark the positions of the most probable breakpoint (the distribution mode). The start and end labels list the offset of the notch from zero and the extent of the confidence intervals. Owing to the abundance of SVs with mapped breakpoints the median resolution of our SV discovery set was 0bp and 10bp for deletions and insertions, respectively.



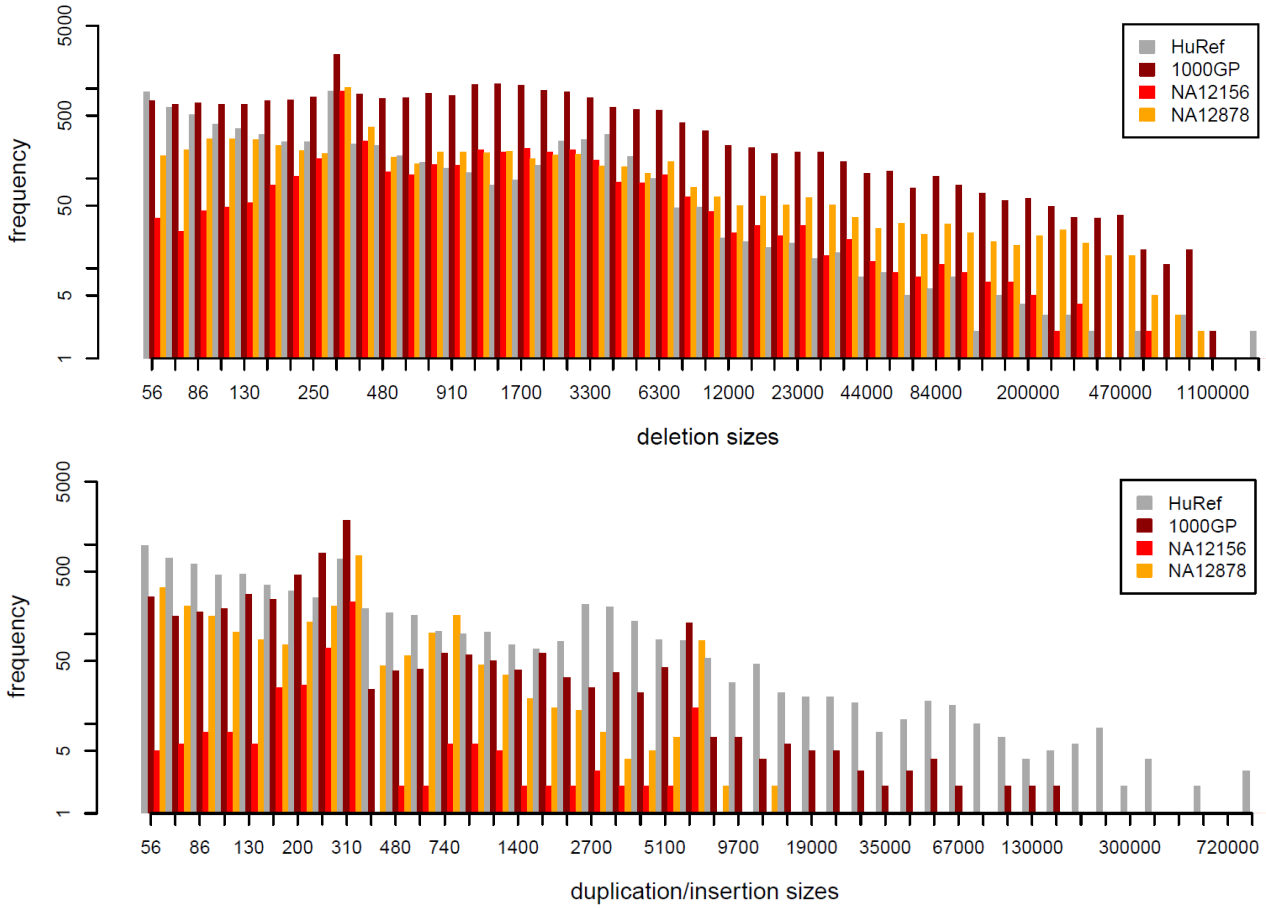
Supplementary Figure 3. Correlation of PCR and array based FDR estimates. FDR estimates based on PCR and arrays are displayed both for trio (blue) and low coverage (red) callsets.



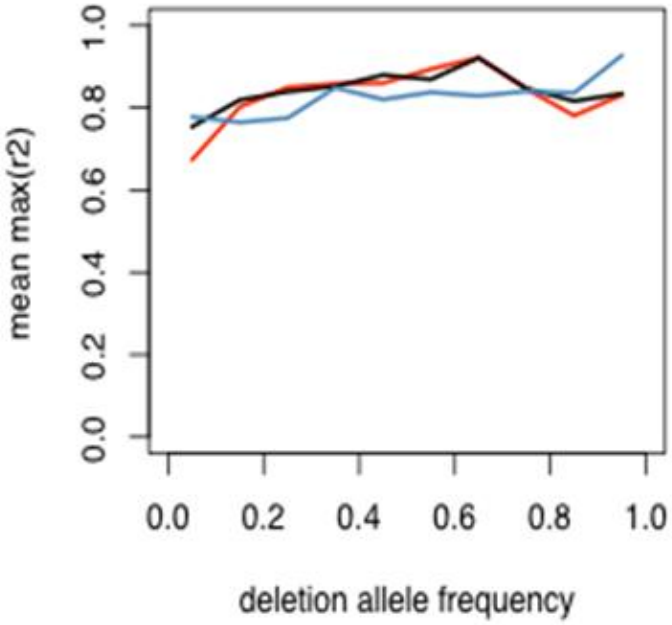
Supplementary Figure 4. Sensitivity of deletion discovery methods for both single and population-scale references. (A) Sensitivity in detecting deletions estimated for three gold standard sources, i.e., sets of deletions identified in capillary read data (median=0.2kb), tiling CGH (2kb), and fosmid sequencing (6kb). Sensitivity estimates are displayed separately for NA12878 (drawn from the trios) and NA12156 (drawn from the low-coverage data). Note that this comparison (based on individual genomes) samples from the universe of variation in a frequency-weighted way. The rightmost panel compares "Rel." (The 1000GP data release) to a set of SVs obtained through the sparse-validation set ("Alg.-centric"). For the low coverage panel, the following methods are displayed: PD=Genome STRiP (using RP and RD features, summarized with 'PD'), RD=Event-wise testing, RP=WTSI RP approach, RL=PEMer, and SR=Pindel. For the trio panel, AS=Cortex, RD=CNVnator, RP=BreakDancer, RL=PEMer, and SR=Yale 454 SR approach. (B) Sensitivity using a population scale reference in low coverage data. For this analysis, the reference data set was based on 3,621 deletions determined by Conrad *et al.*[7] to be polymorphic among the population of genomes analyzed in this study. The sensitivity of each discovery method was calculated as the fraction of these reference deletions for which an intersecting deletion was identified by the method. This and Fig. 2B represent two complementary ways to estimate sensitivity. The reference in Fig. 2B is based on an individual genome (and therefore samples the universe of SV in an allele-frequency-weighted way); the reference here is based on a population of genomes (and therefore contains more low-frequency alleles). Not all methods were applied on all low-coverage genomes (Supplementary Table 1), accounting for some portion of the differences in sensitivity. "Release": release set; "Alg-only": algorithm-centric set.



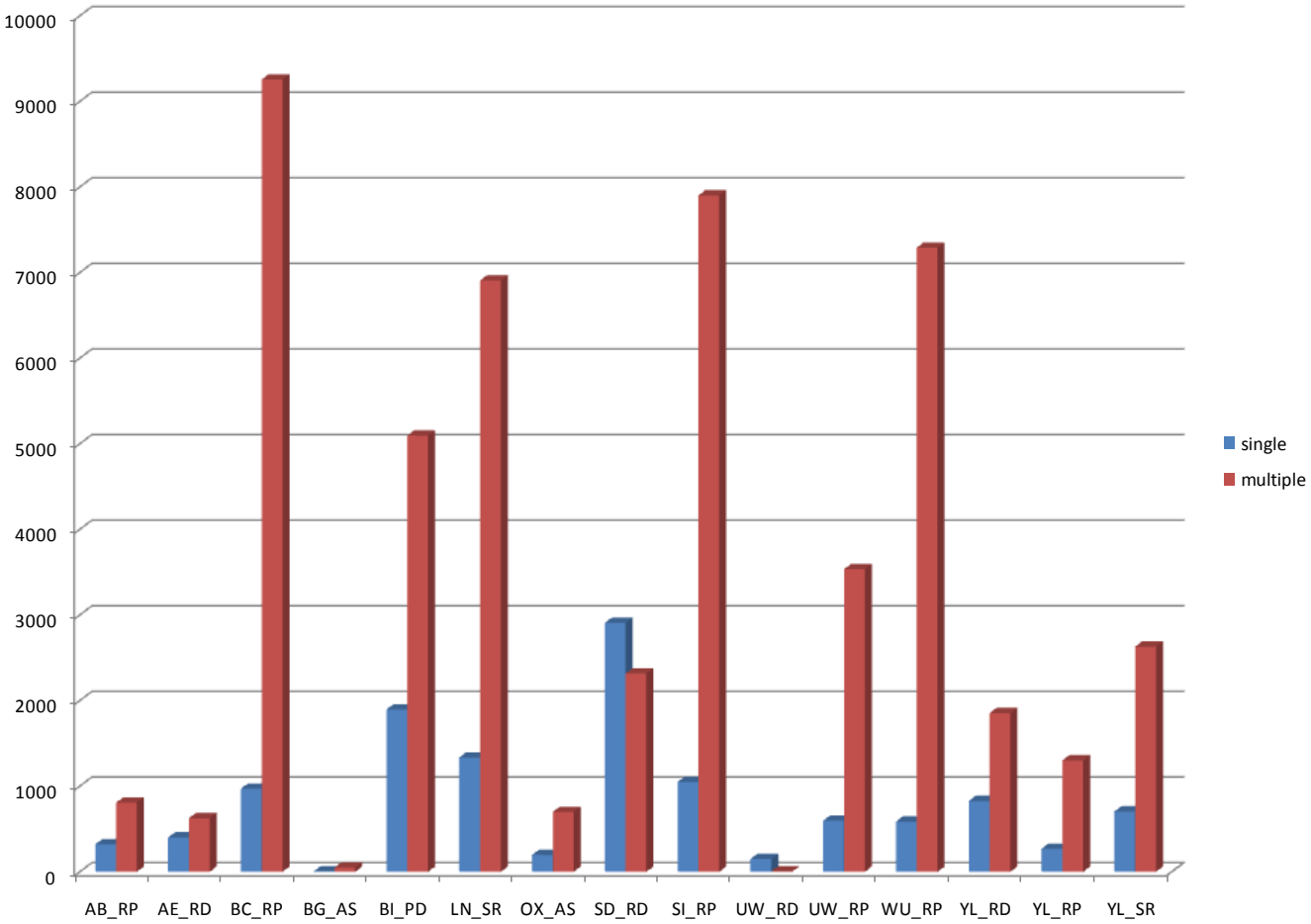
Supplementary Figure 5. Schematic of confidence interval and precision-aware approach for merging SV callsets. Confidence intervals (yellow boxes) around each predicted breakpoint were determined for each callset specific variant (blue line) through comparison with assembled breakpoints (dashed red line). Calls without assembled breakpoints inherited the values derived from each of the respective callset-level assessments. Calls were merged if, and only if, confidence intervals around each breakpoint overlapped. Final breakpoints were assigned using coordinates of assembled breakpoints for a member call (if available) or by using the intersection of the overlapping confidence intervals (vertical red lines).



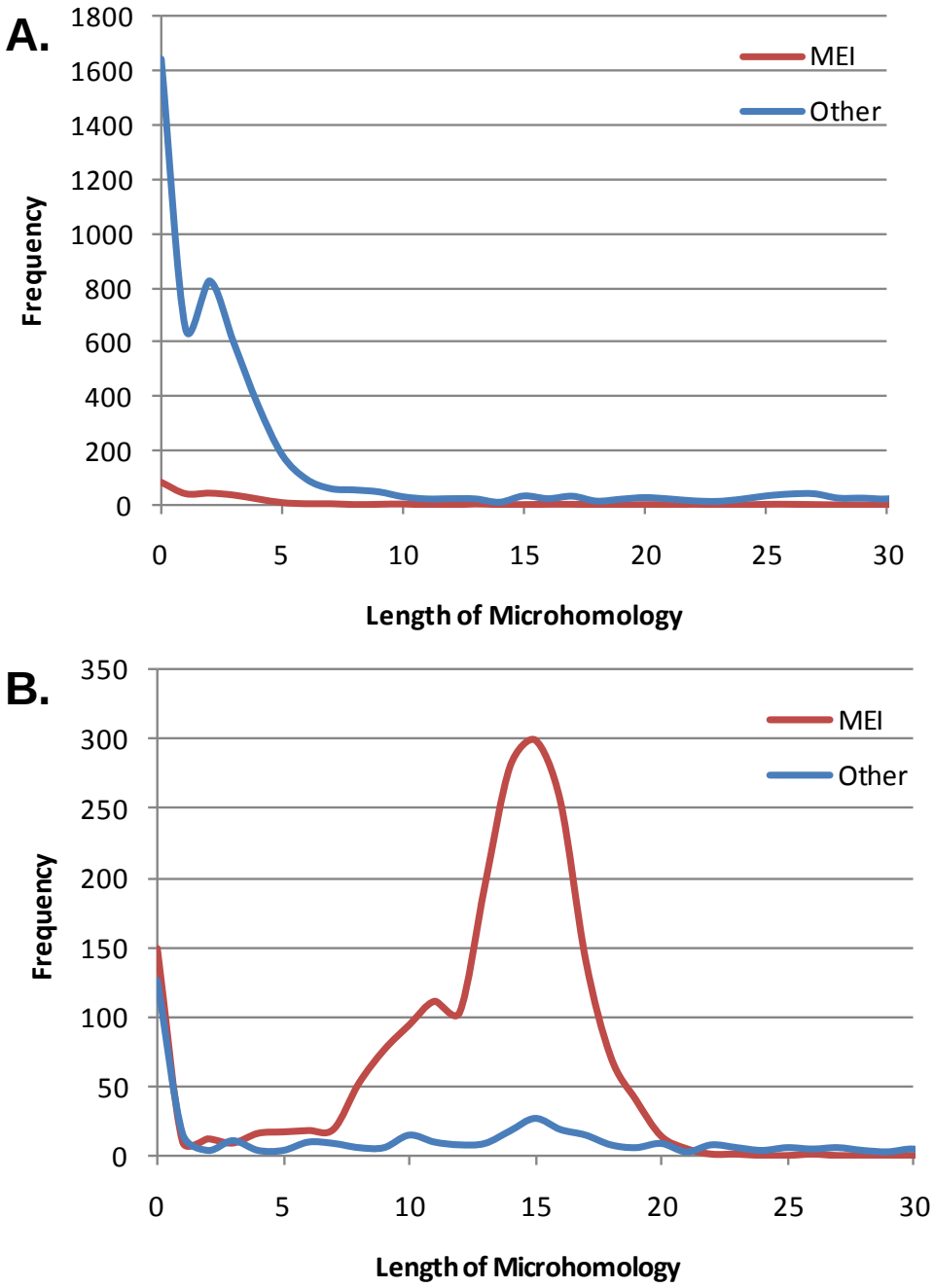
Supplementary Figure 6. Size distribution of discovered SV classes in the Venter genome and our study. Length distributions were compared for both our deletion release set and the combined duplications/insertions set between Pang et al.[11] and our study. The figure also presents the length distributions for two individuals analyzed in our study – NA12156 sequenced at low-coverage and NA12878 sequenced at high-coverage – for direct comparison with the single genome examined in Pang et al. Of note, we identified more large (>5kb) deletions in the release set as well as the trio individual (NA12878) compared to Pang et al., potentially owing to the diversity of SV discovery methods we applied in the trio data and the resultant high sensitivity towards detecting deletions. By comparison, the low-coverage individual (NA12156) appears very similar to Pang et al. in terms of detecting deletions >5kb. Pang et al., however, identified more small deletions (<100bp) than our study; this may in part be due to boundary effects (i.e., some of our discovery methods removed deletions that seemed to be smaller than 50bp (an agreed-upon cutoff) based on a first pass analysis, and did not re-assess those candidates following breakpoint assembly). Furthermore, our study identified overall less insertions/duplications than Pang et al., although Pang et al. and our study appeared to have a similar sensitivity towards Alu (300bp) and LINE (6kb) insertion detection per individual genome.



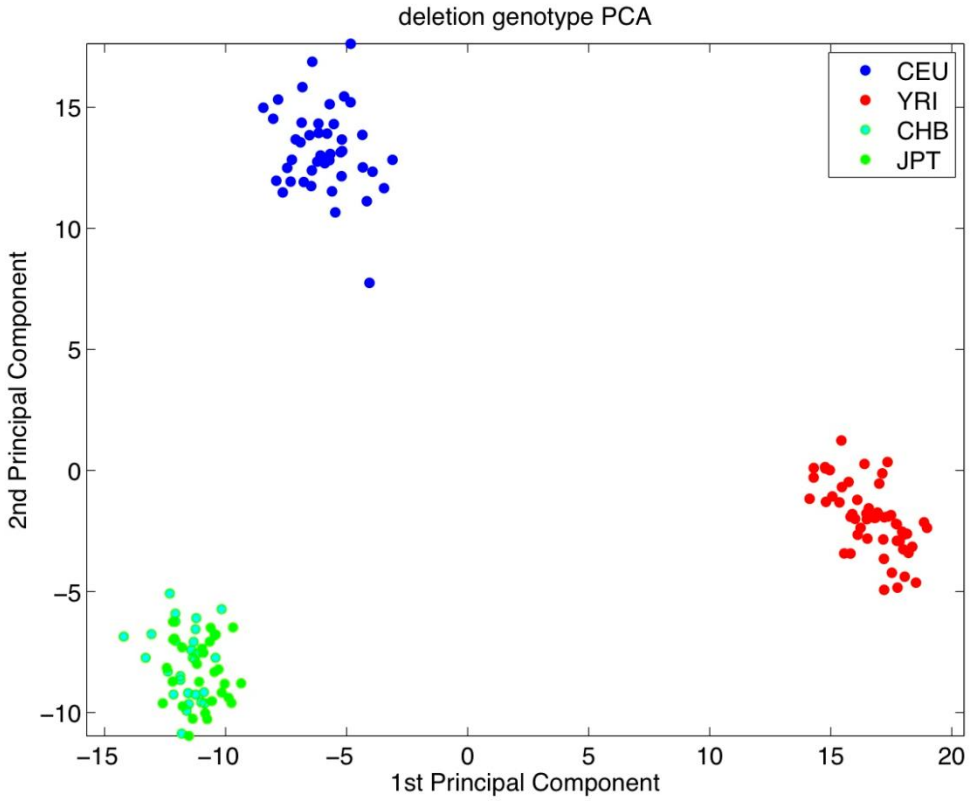
Supplementary Figure 7. Association of deletions genotyped in our study with nearby SNPs. Linkage disequilibrium (LD) between deletions (genotyped in this work) and nearby SNPs (genotyped by HapMap[8]) was assessed by determining the Pearson's correlation coefficient (r^2). For each deletion, the maximum r^2 (among SNPs flanking the breakpoints but within 200kb) was calculated. Data are binned by frequency of the deletion allele and plotted separately for the CEU (red), YRI (blue), and JPT+CHB (black) population samples.



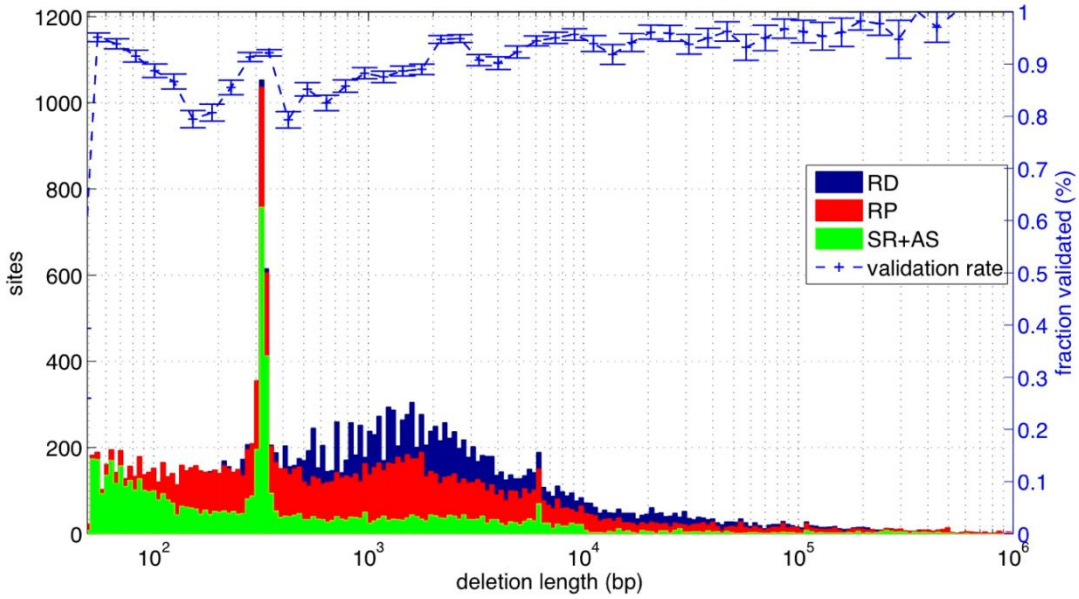
Supplementary Figure 8. Contribution of individual algorithms to deletion union release set. The number of calls from each algorithm was tabulated on the basis of whether each prediction was the only supporting call for a deletion (single) or was one of multiple supporting calls from other algorithms (multiple). SV callsets from discovery methods that were applied to both the trio and low-coverage data were combined.



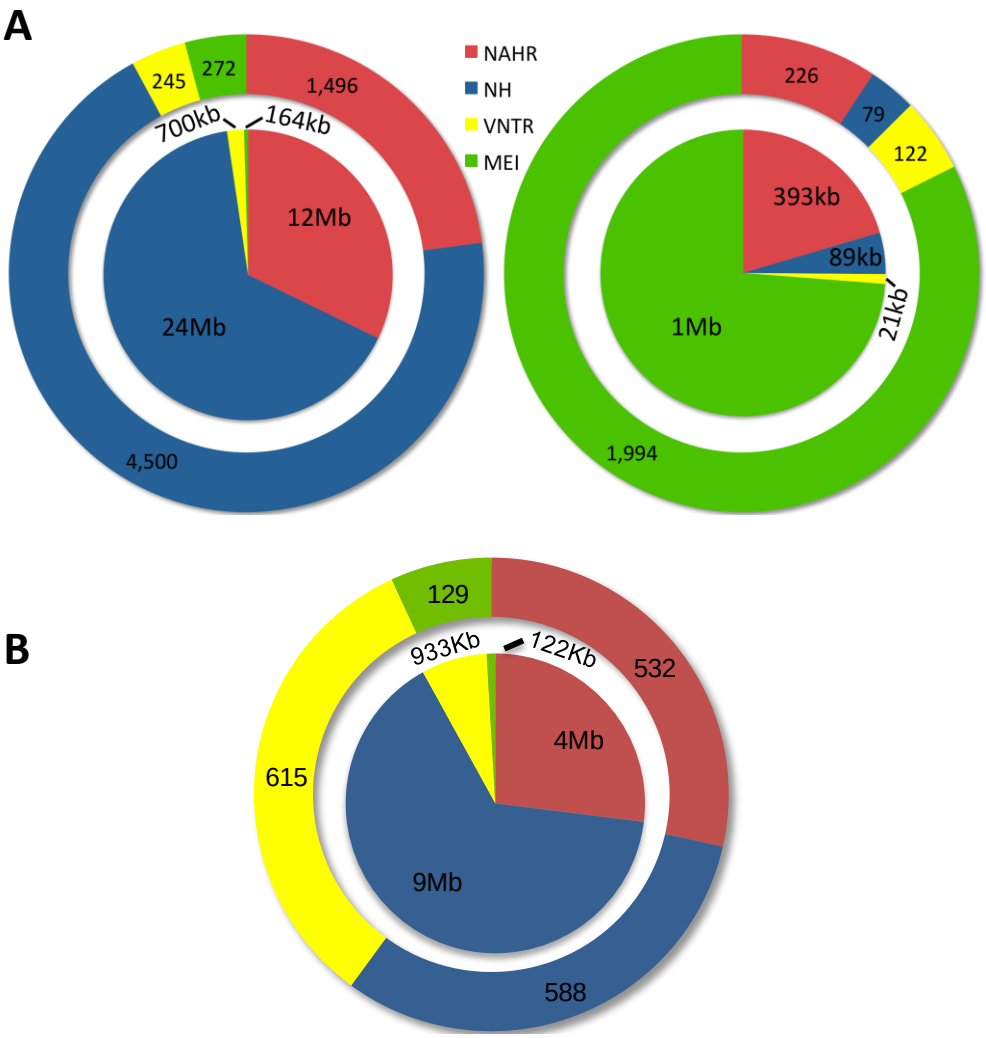
Supplementary Figure 9. Microhomology length distributions from deletions with nucleotide breakpoint resolution. Length distributions are displayed for microhomologies detected at nucleotide resolution breakpoints for (A) deletions and (B) insertions (relative to ancestral status) for events classified as transposable elements (red) or other (blue).



Supplementary Figure 10. Principal component analysis of genotyped deletions.

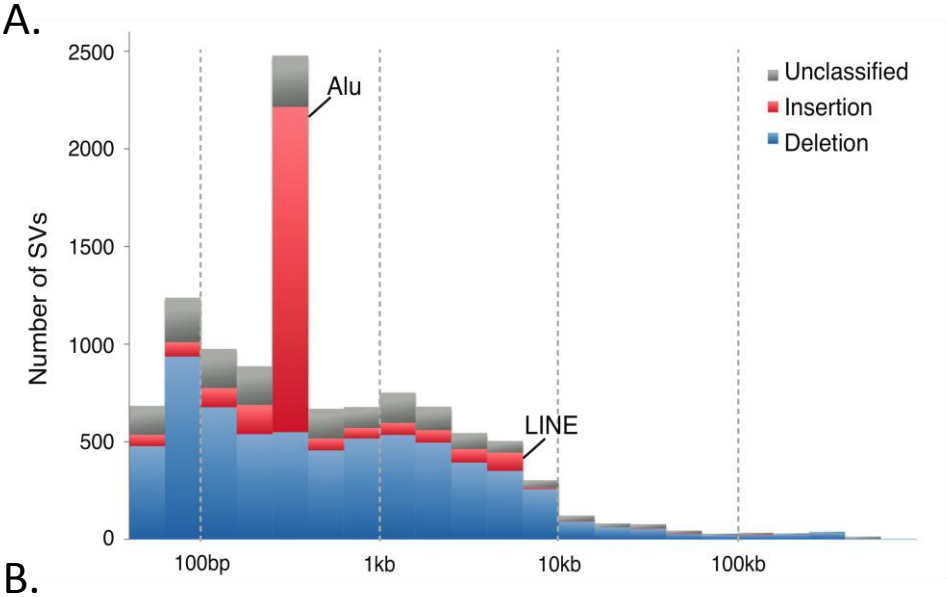


Supplementary Figure 11. Frequency and validation rate of union deletions across the size spectrum. The high validation rate of small events is from the inclusion of primarily validated calls from call sets in that size range. SVs <50bp (leftmost bin in the figure) were excluded from the analyses reported in our study; an analysis of this variation class is reported elsewhere[13].

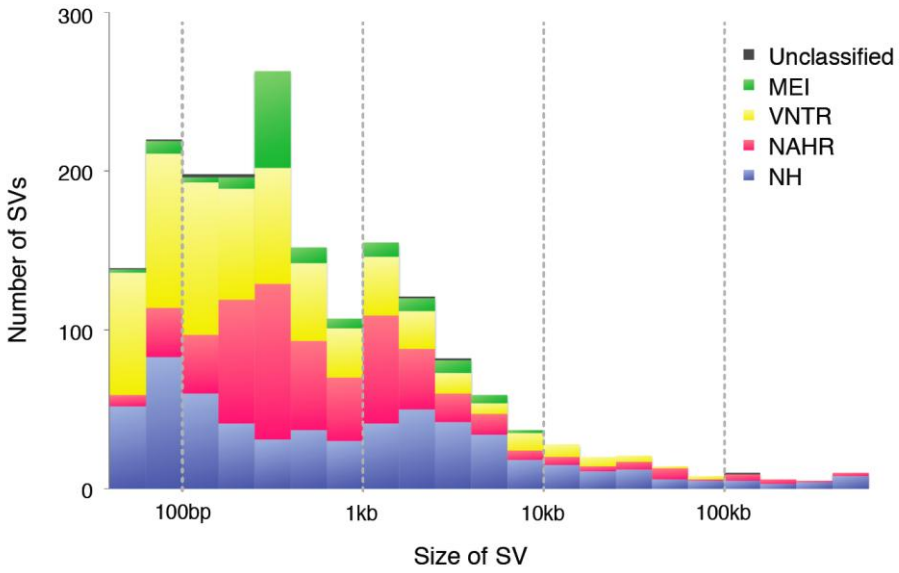


Supplementary Figure 12. Relative contributions of SV formation mechanisms in the genome. (A) Numbers of SVs are displayed on the outer pie chart and affected base pairs on the inner pie chart. Left panel: SVs classified as deletions relative to ancestral loci. Right panel: SVs classified as insertions relative to ancestral loci. (B) Fraction of SVs originally identified as deletions relative to the reference genome, which we were unable to classify with respect to their respective ancestral locus. Numbers of SVs are displayed on the outer pie chart and affected base pairs on the inner pie chart.

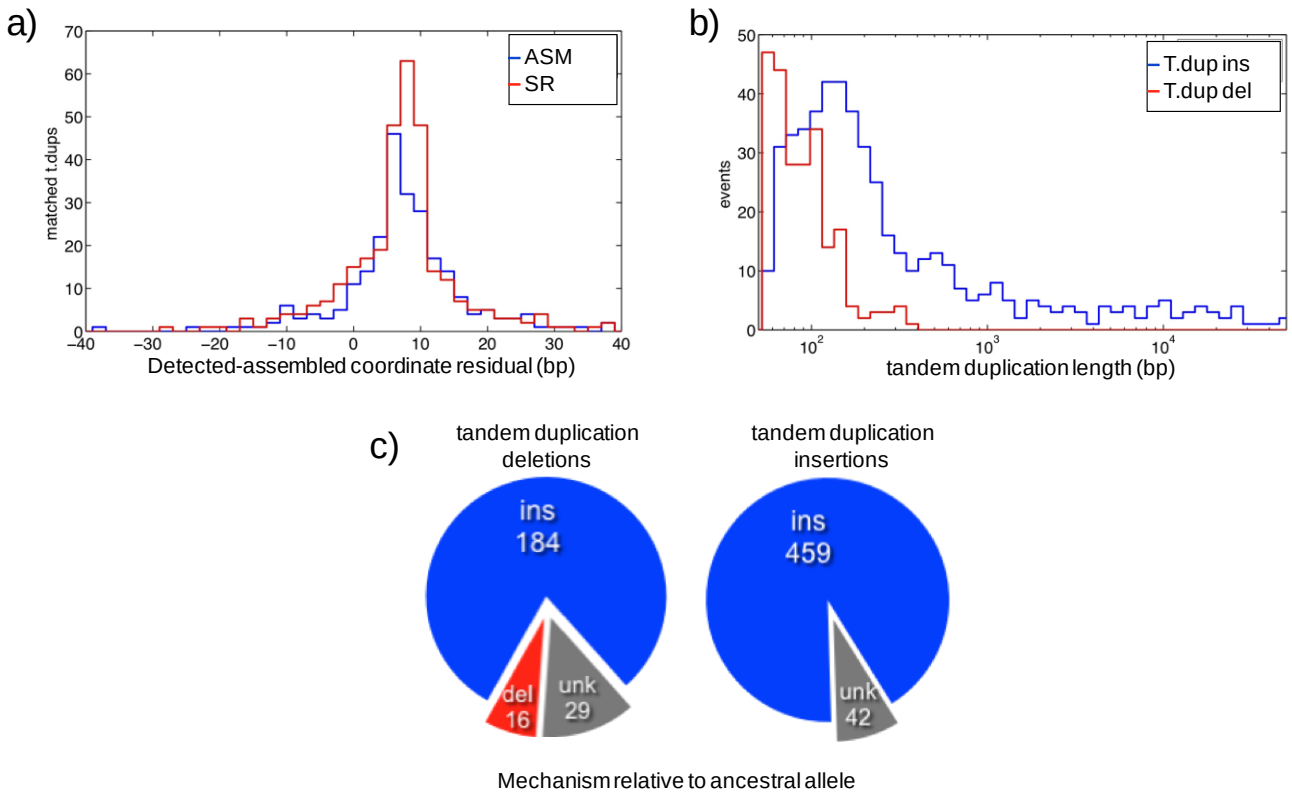
A.



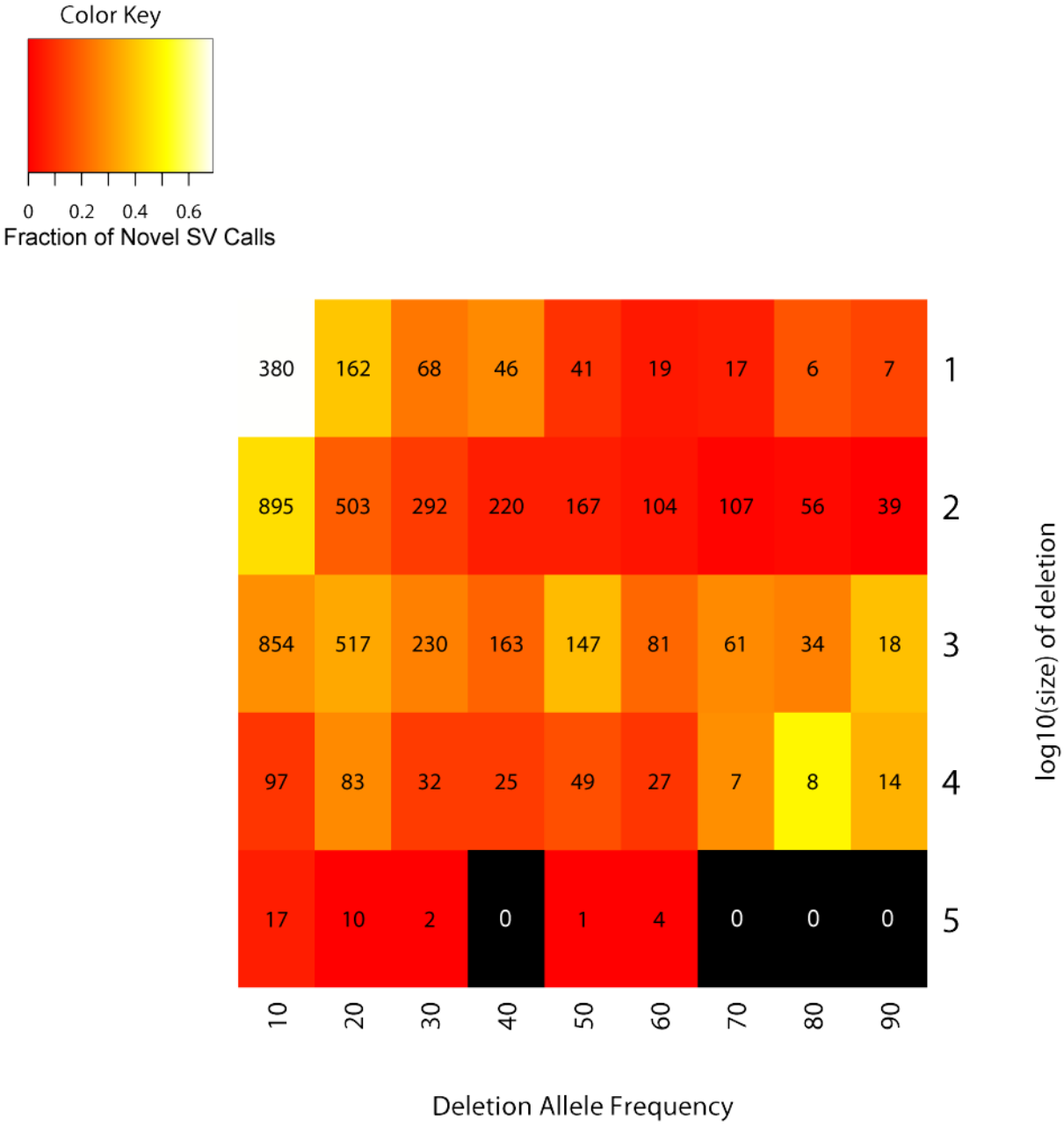
B.



Supplementary Figure 13. **Classification of SVs in terms of ancestral state.** (A) Size spectrum of deletions and insertions relative to ancestral sequence, inferred by aligning deletion breakpoint junctions onto primate genomes. SVs unclassifiable with respect to ancestral state (ambiguous ancestral state classification) are displayed in gray. (B) Formation mechanism size spectrum for SVs unclassified in terms of ancestral state (SVs unclassifiable with respect to formation mechanism are displayed in black.)



Supplementary Figure 14. Tandem duplication properties. a) Breakpoint coordinate resolution for tandem duplications detected as insertions. RP detected events were matched by 50% reciprocal overlap to local assembled[14] tandem duplications (ASM) and split-read (SR) in the same data. RP event start and end residuals have a standard deviation of 10bp and an offset from 0 of 10bp to both ASM and SR positions. **b)** Tandem duplication lengths for events detected as deletions (red) and insertions (blue). Length sensitivity for deletions extends from 50bp to 300bp while insertion detection extends beyond 10kb. **c)** Pie chart of ancestral allele classifications for tandem duplications; 'ins' refers to insertions relative to the ancestral allele, 'del' to deletions, and 'unk' unknown. Both types are dominated by insertions of the duplication relative to the ancestor genome.



Supplementary Figure 15. Fraction of novelty stratified by SV size and allelic frequency. Heatmap colors range from a low fraction observed to be novel (darker) to a high fraction (lighter). Black signifies no data for a particular bin. The numbers in each square indicate the total number of deletions falling into a particular size and frequency bin.

Supplementary Table 1. Sequencing statistics for SV discovery. For each sample, the median insert size was calculated by taking the mean of the median insert sizes of all sequencing files of this sample for either low coverage or trio data ("NA" indicates libraries sequenced by single ended sequencing). Sequence coverage was determined by taking the sum of the number of mapped bases in all alignment files divided by the accessible genome size (2.4G for low coverage and 2.3G for trio). For paired end sequencing, physical coverage was determined based on the number of mapped read-pairs multiplied with the median insert size and divided by the accessible genome size. (Physical coverage was set to sequence coverage for single ended sequencing.)

External File

Supplementary Table 2A. SV discovery sets in low-coverage sequencing data

Approach	Callset Origin	Discovery Algorithm Name and Reference*	Platform	Mapping Algorithm	Genomes Analyzed	SV Type	Algorithm Parameters Used
RD	**AE**	**N/A**[13]	Illumina	MAQ	8	DEL	window size (≥500bp); p-value ($P \leq 10^{-5}$)
	SD	**Event-wise testing**[15,#]	Illumina	MAQ	162	DEL	read mapping quality (≥Q30); window size (100bp); cluster size with merged events of same type(≤ 500bp); read depth (≤0.75 and ≥ 1.25 mean read depth); significance level ($P<10^{-6}$); event size (≥ 1kb); absolute difference between median read counts(>0.5)
	YL	**CNVnator**[10]	Illumina	MAQ	65	DEL	N/A
RP	**BC**	**Spanner**[13]	Illumina	MOSAIK	138	DEL	maximum mismatch threshold (4 for 36-43mer reads, 6 for 44-63mers, and 12 for 64mers and longer); hash size (15); Smith-Waterman bandwidth (17); alignment candidate threshold (25bp); local alignment search radius (100bp); hash position threshold (100); mapping distance (P-value≥0.99); minimum read-pairs (4, 2 from each side); map distance to annotated loci (≥400bp); gap between the F and R clusters (-30 bp < gap < 500 bp)
	BC	**Spanner**[13]	Illumina	MOSAIK	138	INS	maximum mismatch threshold (4 for 36-43mer reads, 6 for 44-63mers, and 12 for 64mers and longer); hash size (15); Smith-Waterman bandwidth (17); alignment candidate threshold (25bp); local alignment search radius (100bp); hash position threshold (100)
	SI	**N/A**[13]	Illumina	MAQ	144	DEL	MAQ mapping quality (≥20); read-pairs in a cluster (≥2); start/end distance (10 x median absolute deviation of the insert size distribution); event size (<1Mb)
	YL	**PEMer**[16,23]	SOLiD	CORONA	25	DEL	span-size (within 15% deviation from the median of span-size)
	WU	**BreakDancer**	Illumina	MAQ	138	DEL	RMAQ mapping quality (> 35); outer distance (> mean + 4stdev of the insert size)
SR	**BC**	**Mosaik**[13]	454	MOSAIK	22	INS	hash size (15bp); mismatch bases in alignments (≤5%); match bases aligned to one of the mobile element consensus sequences (40bp); gap length (≤6bp); alignment quality score (≥40);mobile element alignment length (>60bp); distance from annotated mobile elements (≥100bp)
	LN	**Pindel**[17]	Illumina	MAQ	145	DEL	MAQ mapping quality (>0); maximum deletion size (50kb); number of fragments for unmapped reads (2 for deletion and 3 for short insertions)
	LN	**Pindel**[17]	Illumina	MAQ	145	INS	MAQ mapping quality (>0); maximum deletion size (50kb); number of fragments for unmapped reads (2 for deletion and 3 for short insertions)
	YL	**N/A**[13]	454	BLAT	5	DEL	N/A
	YL	**N/A**[13]	454	BLAT	5	INS	N/A
PD	**BC**	**Spanner**[13]	Illumina	MOSAIK	138	TDUP	mapping quality values of read pairs (≥30); mapping distance between the pairs (p-value<0.02%); number of supporting read pairs (≥3); minimum deletion size (50bp); "Alignability" in the clustered regions (> 0.01); Net read c overage over all samples (< 2.5 x the expected coverage); event length (≤250bp); copy number (≤2.2)
	BI	**Genome STRiP**[18]	Illumina	MAQ	168	DEL	clusters of paired-ends (N >= 2); apparent insert size (> the median of the insert size distribution + 10 x the median absolute deviation of insert size from the median for that lane/library);

*unpublished algorithms are described in detail in the Supplementary Information of the main 1000GP paper[13]
[#]uses a population scale reference made up from multiple low coverage samples
Of note, not all SV discovery methods were applied across all genomes. For example, RP methods using the long (>1kb) insert sizes from the SOLiD or 454 platforms (abbreviated as RL in the main text) were applied to genomes for which such long insert size libraries were available.

Supplementary Table 2B. SV discovery sets in trio sequencing data

Approach	Callset Origin	Discovery Algorithm Name and Reference*	Platform	Mapping Algorithm	Genomes analyzed	SV Type	Algorithm Parameters Used
RD	**SD**	**Event-wise testing**[15]	Illumina	MAQ	6	DEL	read mapping quality (≥Q30); window size (100bp); cluster size with merged events of same type(≤ 500bp); read depth (≤0.75 and ≥ 1.25 mean read depth); significance level (P<10-6); event size (≥ 1kb); absolute difference between median read counts (>0.5); median read-depth (<1.25); common deletion regions (>4 occurrencies)
	UW	**mrFAST**[19]	Illumina	mrFAST	6	DEL	RepeatMasker (on human reference genome build 35, with the sensitivity option "-s" enabled); Tandem Repeats Finder (mask tandem repeats ≤500bp); edit distance (≤ 2); unique PDervals (5 kb of unmasked sequence); windows (6/7 consecutive 5 kb windows with read depth >average-2stdev)
	YL	**CNVnator**[10]	Illumina	MAQ	6	DEL	N/A
RP	**AB**	**AB large indel tool**[13]	SOLiD	MAPREADS	1	DEL	read-pairs in a cluster (≥2)
	BC	**Spanner**[13]	Illumina	MOSAIK	6	DEL	maximum mismatch threshold (4 for 36-43mer reads, 6 for 44-63mers, and 12 for 64mers and longer); hash size (15); Smith-Waterman bandwidth (17); alignment candidate threshold (25bp); local alignment search radius (100bp); hash position threshold (100); mapping distance (P-value≥0.99); minimum read-pairs (4, 2 from each side); map distance to annotated loci (≥400bp); gap between the F and R clusters (-30 bp < gap < 500 bp)
	BC	**Spanner**[13]	Illumina	MOSAIK	6	INS	maximum mismatch threshold (4 for 36-43mer reads, 6 for 44-63mers, and 12 for 64mers and longer); hash size (15); Smith-Waterman bandwidth (17); alignment candidate threshold (25bp); local alignment search radius (100bp); hash position threshold (100)
	SI	**N/A**[13]	Illumina	MAQ	6	DEL	MAQ mapping quality (≥20); read-pairs in a cluster (≥2); start/end distance (10 x median absolute deviation of the insert size distribution); event size (<1Mb)
	UW	**Variation Hunter**[20]	Illumina	mrFAST	6	DEL	high-quality reads (average phred score ≥ 20); edit distance (≤ 2 with the mrFAST); size threshold (average± 4×stdev)
	WU	**BreakDancer**	Illumina	MAQ	6	DEL	MAQ mapping quality (> 35); outer distance (> mean + 4stdev of the insert size)
	YL	**PEMer**[16,23]	454	PEM	1	DEL	p-value cutoff of 0.05
	YL	**PEMer**[16,23]	454	PEM	1	INS	p-value cutoff of 0.05
SR	**BC**	**Mosaik**[13]	454	MOSAIK	2	INS	hash size (15bp); mismatch bases in alignments (≤5%); match bases aligned to one of the mobile element consensus sequences (40bp); gap length (≤6bp); alignment quality score (≥40);mobile element alignment length (>60bp); distance from annotated mobile elements (≥100bp)
	LN	**Pindel**[17]	Illumina	MAQ	6	DEL	MAQ mapping quality (>0); maximum deletion size (50kb); number of fragments for unmapped reads (2 for deletion and 3 for short insertions)
	YL	**N/A**[13]	454	BLAT	1	DEL	N/A
	YL	**N/A**[13]	454	BLAT	1	INS	N/A
AS	**BG**	**SOAPdenovo**[21]	Illumina	SOAP	6	DEL	prealignment (BLAT v. 30 with –fastMap and –maxPDron=50); scaffold set alignment (LASTZ V1.01.50 with high-scoring segment pairs (HSP) chaining option, ambiguous 'N' treatment, and gap-free extension tolerance up to 50kb); Best hits were further confirmed using "axtBest"
	BG	**SOAPdenovo**[21]	Illumina	SOAP	6	INS	prealignment (BLAT v. 30 with –fastMap and –maxPDron=50); scaffold set alignment (LASTZ V1.01.50 with high-scoring segment pairs (HSP) chaining option, ambiguous 'N' treatment, and gap-free extension tolerance up to 50kb); Best hits were further confirmed using "axtBest"
	OX	**Cortex**[13]	Illumina	CORTEX	1	DEL	event size (≤1kb for "bubble calling" algorithm and ≤40kb for "reference assisted" algorithm)
	OX	**Cortex**[13]	Illumina	CORTEX	1	INS	event size (≤1kb for "bubble calling" algorithm and ≤40kb for "reference assisted" algorithm)
	UW	**NovelSeq**[22]	Illumina	mrFAST	6	INS	event size (≥200bp)
PD	**BC**	**Spanner**[13]	Illumina	MOSAIK	6	TDUP	mapping quality values of read pairs (≥30); mapping distance between the pairs (p-value<0.04%); number of supporting read pairs (≥3); minimum deletion size (50bp); "Alignability" in the clustered regions (> 0.01); Net read coverage over all samples (< 2.5 x the expected coverage); event lenght (≤250bp); copy number (≤2.2)

*unpublished algorithms are described in detail in the Supplementary Information of the main 1000GP paper[13]
Of note, not all SV discovery methods were applied across all genomes. For example, RP methods using the long (>1kb) insert sizes from the SOLiD or 454 platforms (abbreviated as RL in the main text) were applied to genomes for which such long insert size libraries were available.

Supplementary Table 3. Complete list of low coverage calls by institution and set. This table contains the raw, unfiltered output from for each different discovery algorithms applied to the low coverage data set, as well as a summary tab describing various statistics. SVs <50bp were neither assessed by experimental validation nor included in our "release set".

External File

Supplementary Table 4. Complete list of trio calls by institution and set. This table contains the raw, unfiltered output from for each different discovery algorithms applied to the trio data set, as well as a summary tab describing various statistics. SVs <50bp were neither assessed by experimental validation nor included in our “release set”.

External File

Supplementary Table 5. Gold standard SV sets for NA12878 and NA12156 from 4 external and orthogonal data set

External File

Supplementary Table 6A. Sensitivity in discovering deletions for different methods, assessed in NA12156(*)

Approach	Callset Origin	Algorithm	Sequencing platform	Kidd (n=54)	Conrad (n=353)	McCarroll (n=118)	Mills (n=151)
RD	**SD**	**Event-wise testing**	Illumina	0.46	0.65	0.70	0.06
	YL	**CNVnator**	Illumina	0.20	0.19	0.31	0.09
RP	**BC**	**Spanner**	Illumina	0.26	0.19	0.17	0.21
	SI	**N/A**	Illumina	0.30	0.28	0.25	0.21
	YL	**PEMer**	SOLiD	0.11	0.28	0.09	0.03
	WU	**BreakDancer**	Illumina	0.20	0.20	0.18	0.17
	LN	**Pindel**	Illumina	0.13	0.08	0.13	0.10
PD	**BI**	**Genome STRiP**	Illumina	0.63	0.50	0.40	0.21

(*) Methods not used for SV discovery in this individual are not shown

Supplementary Table 6B. Sensitivity in discovering deletions for different methods, assessed in NA12878(*)

Approach	Callset Origin	Algorithm name	Sequencing platform	Kidd (n=58)	Conrad (n=373)	McCarroll (n=130)	Mills (n=81)
RD	**SD**	**Event-wise testing**	Illumina	0.67	0.56	0.80	0.05
	UW	**mrFAST**	Illumina	0.16	0.07	0.22	0.00
	YL	**CNVnator**	Illumina	0.91	0.84	0.88	0.24
RP	**BC**	**Spanner**	Illumina	0.45	0.50	0.32	0.44
	SI	**N/A**	Illumina	0.50	0.55	0.42	0.24
	UW	**VariationHunter**	Illumina	0.55	0.53	0.50	0.30
	WU	**BreakDancer**	Illumina	0.50	0.55	0.44	0.40
	YL	**PEMer**	454	0.91	0.45	0.72	0.10
SR	**LN**	**Pindel**	Illumina	0.28	0.38	0.25	0.28
	YL	**N/A**	454	0.55	0.54	0.44	0.52
AS	**BG**	**SOAPdenovo**	Illumina	0.03	0.00	0.01	0.01
	OX	**Cortex**	Illumina	0.07	0.04	0.13	0.20

(*) Methods not used for SV discovery in this individual are not shown

Supplementary Table 7. Summary of the SV Release Set and of the Algorithm-centric set. "Single, FDR<10%" refers to SVs from individual callsets that met the predefined FDR threshold. The "validated" set includes the number of total calls validated for each data set. "Pairwise, FDR<10%" refers to SV calls meeting the predefined FDR threshold which were identified by a pairwise integration of distinct SV discovery methods (on the basis of callset intersection). The "release set" is made up of all non-redundant calls falling in the "single, FDR<10%" or "validated" sets. The "algorithm-centric set" is made up of all non-redundant calls falling in the "single, FDR<10%" or "pairwise, FDR<10%" sets. Note that the SV sets "single, FDR<10%", "validated", and "pairwise, FDR<10%" are obviously not mutually exclusive, *i.e.*, some SVs fall into all three sets.

SV class	set	low-coverage	trio	union
deletions	single, FDR<10%	8906	6360	11215
	validated	14576	9695	19767
	pairwise, FDR<10%	6603	5447	8458
	release set	15893	11248	22025
	algorithm-centric set	9567 (60%)	7336 (65%)	12642 (57%)
tandem duplications	single, FDR<10%	407	256	501
	validated	74	88	88
	pairwise, FDR<10%	0	0	0
	release set	407	256	501
	algorithm-centric set	407 (100%)	256 (100%)	501 (100%)
mobile element insertions	single, FDR<10%	4775	2531	5371
	validated	796	724	870
	pairwise, FDR<10%	1688	1375	1793
	release set	4775	2531	5371 (100%)
	algorithm-centric set	4775 (100%)	2531 (100%)	5371 (100%)
novel insertions	single, FDR<10%	-	-	-
	validated	-	128	128
	pairwise, FDR<10%	-	-	-
	release set	-	128	128
	algorithm-centric set	-	-	-
Total	**release set**	**20775**	**14163**	**28025**
	algorithm-centric set	**14749 (71%)**	**10123 (71%)**	**18514 (67%)**

(*) Numbers in parenthesis indicate fraction of release set captured by algorithm-centric set

Supplementary Table 8A. SV discovery method sensitivity in low-coverage data, assessed based on the combined gold standard set derived from individual NA12156 using different overlap criteria

Approach	Callset Origin	Any overlap	50% reciprocal overlap
RD	**SD**	0.508	0.297
	YL	0.190	0.142
RP	**BC**	0.199	0.185
	SI	0.259	0.229
	YL	0.172	0.09
	WU	0.188	0.164
SR	**LN**	0.099	0.088
PD	**BI**	0.426	0.372
	release-set	0.694	0.505
	algorithm-centric set	0.578	0.564

Supplementary Table 8B. SV discovery method sensitivity in trio data, assessed based on the combined gold standard set derived from individual NA12878 using different overlap criteria

Approach	Callset Origin	Any overlap	50% reciprocal overlap
RD	**SD**	0.556	0.444
	UW	0.097	0.042
	YL	0.775	0.634
RP	**SI**	0.475	0.424
	UW	0.494	0.420
	WU	0.499	0.432
	YL	0.504	0.401
	BC	0.452	0.406
SR	**LN**	0.333	0.295
	YL	0.517	0.395
AS	**BG**	0.006	0.006
	OX	0.082	0.048
	release-set	0.816	0.697
	algorithm-centric set	0.841	0.720

Supplementary Table 9. Functional analysis of deletions which overlap transcripts. The set of CNV deletions was intersected with the Gencode v3 set of transcripts to assess their possible functional impact. For positive strand mRNAs, the CNV start and end positions correspond to left and right boundaries of a CNV deletion, respectively. For negative strand mRNAs, the start and end positions are reversed. "Functional_regions" are classified according to their biological function as following (from the upstream to the downstream region of a transcript): upstream, promoter, 5'UTR, start_codon, CDS, stop_codon, 3'UTR, downstream. Note that start_codon and stop_codon are listed separately - the "CDS" category does not include start/stop codon. The length of promoter is set to 200 bp. "Splicing regions" are classified as "Exon" and "Non-exon", with "Exon" corresponding to exonic regions of the transcript that overlap with CNV deletions, and "Non-exon" corresponding to intronic and out-of-transcripts boundary regions that overlap with CNV deletions. "Function_change" categories includes: Regulatory_change (changes that only influence transcription and/or translation); Transcriptional_change (changes that completely abolish transcription); Coding_change (changes that influence encoded protein sequence or abolish protein production); and No_effect (no changes in promoter or exons). "Type" categories include: In_frame (no changes of reading frame); Out_of_frame (changes of reading frame); Truncation (truncation of the coding sequence); Elongation (elongation of the coding sequence due to Stop codon deletion); Loss (elimination of the entire transcript or encoded protein product); Promoter_interrupted, 5'UTR_interrupted, 3'UTR_interrupted (removal of the parts of corresponding regions); Splice_site_deleted (removal of the part(s) of exon(s)); NA (not available).

External File

Supplementary Table 10. Gene Ontology (GO) enrichment analysis for deletions overlapping protein coding regions. Gene ontology (GO) enrichment was performed using GOstat (Reference to GOstat) that searches for statistically under- and over-represented GO terms within the protein coding regions of deletions. Benjamini multiple testing was used and a p-value cutoff was set at 0.05. Those GO terms with hierarchy levels of 3 or higher for biological processes and molecular functions are recorded.

External File

Supplementary Table 11. Formation mechanisms and ancestral states of SVs inferred with the BreakSeq pipeline. NAHR: non-allelic homologous recombination events mediated by >50bp homologous sequences of >85% identity at the two breakpoints. VNTR: variable number of tandem repeats, where >50% of the SV region is annotated as simple repeats, satellite repeats, or low-complexity regions. MEI: mobile element insertion associated events, consistent with one or multiple transposable element insertion, mostly long and short interspersed elements (LINEs and SINEs) and combinations thereof; analysis includes assessment of target site duplications and a poly A tract at the 3' end. NH: non-homologous events, including non-homologous end-joining (NHEJ) and replication fork collapse-associated (FoSTeS/MMBIR). NAHR and MEI have a high-confident subset (if the corresponding sequence signatures are all present) and an extended subset of medium confidence (if some of the corresponding sequence signatures are missing) respectively. The extended subsets have a _EXT suffix. All our analyses combined the higher and medium confidence callsets (trends reported in our paper were robust towards excluding the medium confident callsets).

External File

Supplementary Table 12. Formation mechanism inference with BreakSeq for deletions identified with different SV discovery methods. These data were generated using SVs that either were validated by local targeted assembly (with TIGRA) or that by themselves represented experimentally validated nucleotide resolution calls (e.g. SR based calls). We formed the non-redundant union of calls made in low-coverage and trio data. Note that apart from the types of approaches used also the sequencing platform affected the propensity for detecting SVs from a certain formation mechanism class. For example, while Pindel was the SR method detecting most SVs, the Yale SR method operating with long (Roche) DNA reads detected more SVs associated with repetitive sequence (VNTR and NAHR).

Approach	Callset Origin	Algorithm	VNTR	NAHR	NH	MEI	MEI-assoc,	Unclassified	Total number of events
RD	**UW**	**mrFAST**	0.0%	50.0%	50.0%	0.0%	0.0%	0.0%	2
	YL	**CNVnator**	5.1%	8.5%	74.6%	10.2%	1.7%	0.0%	59
RP	**BC**	**Spanner**	3.9%	10.3%	45.6%	36.9%	2.4%	0.8%	5241
	YL	**PEMer**	0.5%	10.4%	59.9%	18.3%	7.4%	3.5%	202
	UW	**VariationHunter**	3.4%	10.1%	33.3%	49.7%	2.3%	1.2%	3198
	WU	**BreakDancer**	5.7%	14.3%	38.2%	38.1%	2.4%	1.3%	4879
	SI	**N/A**	2.4%	13.6%	43.1%	37.0%	2.8%	1.2%	5087
SR	**LN**	**Pindel**	4.7%	1.8%	58.7%	31.7%	2.1%	1.1%	6439
	YL	**N/A**	15.8%	20.2%	26.6%	34.2%	1.6%	1.5%	4203
AS	**OX**	**Cortex**	7.0%	25.2%	15.2%	50.5%	1.6%	0.6%	703
PD	**BI**	**Genome STRiP**	1.1%	10.5%	57.8%	26.8%	2.8%	1.0%	3995
Whole deletion set	**-**	**-**	10.7%	21.0%	43.9%	21.9%	1.7%	0.9%	13159

Supplementary Table 13. Enrichment of discovered union SVs near recombination hotspots and segmental duplication

Type	Mechanism	No. of Breakpoints	No. of overlapping Breakpoints	Expected	Enrichment	P-value	Permutation P-value
Recombination Hotspots	**NAHR**	6300	682	504.0	1.35	1.33E-15	5.16E-08
	NH	12238	1022	979.0	1.04	7.45E-02	4.52E-01
	MEI	7022	633	561.8	1.13	9.56E-04	6.32E-02
	VNTR	2970	262	237.6	1.10	4.77E-02	2.19E-01
Segmental Duplications	**NAHR**	6300	469	315.0	1.49	3.11E-17	4.31E-02
	NH	12238	573	611.9	0.94	9.45E-01	1.08E-02
	MEI	7022	155	351.1	0.44	9.97E-01	2.50E-07
	VNTR	2970	322	148.5	2.17	1.00E-37	4.26E-04

Supplementary Table 14. List of identified putative mechanistic hotspots

Chr	Start (hg18)	End (hg18)	Enrich-ment	Adj. *P*-value	Composition of hotspot in terms of formation mechanisms (portion in parentheses)	Overlap of hotspot with region associated with disorder (according to DECIPHER[23])
1	555671	1256457	5	5.5E-03	NAHR(0.33),VNTR(0.11),NH(0.56),MEI(0.00)	1p36 microdeletion syndrome (chr1:1-5,308,621)
1	234268582	234768582	5	5.5E-03	NAHR(0.33),VNTR(0.33),NH(0.11),MEI(0.22)	
1	243844783	244344783	5	5.5E-03	NAHR(0.56),VNTR(0.22),NH(0.22),MEI(0.00)	
10	576921	1092002	5	5.5E-03	NAHR(0.80),VNTR(0.20),NH(0.00),MEI(0.00)	
10	3080737	3580737	5	1.6E-03	NAHR(0.56),VNTR(0.22),NH(0.22),MEI(0.00)	
10	41678275	42213642	6	5.5E-03	NAHR(0.00),VNTR(0.92),NH(0.08),MEI(0.00)	
10	133602620	135298931	7	1.1E-04	NAHR(0.82),VNTR(0.13),NH(0.05),MEI(0.00)	
11	180951	880479	5	1.5E-05	NAHR(0.80),VNTR(0.10),NH(0.00),MEI(0.10)	
11	48304178	48808207	5	1.1E-04	NAHR(0.09),VNTR(0.36),NH(0.27),MEI(0.27)	
11	133864554	134364554	5	1.6E-03	NAHR(0.67),VNTR(0.11),NH(0.22),MEI(0.00)	
12	34247220	34906743	5	1.6E-03	NAHR(0.00),VNTR(0.70),NH(0.30),MEI(0.00)	
13	20384068	21078351	5	5.5E-03	NAHR(0.20),VNTR(0.10),NH(0.70),MEI(0.00)	
13	112324851	113291477	5	1.6E-03	NAHR(0.78),VNTR(0.06),NH(0.17),MEI(0.00)	
13	113565279	114107303	6	5.5E-03	NAHR(0.64),VNTR(0.09),NH(0.27),MEI(0.00)	
14	21826293	22326293	5	1.6E-03	NAHR(0.22),VNTR(0.00),NH(0.56),MEI(0.22)	
14	105269503	105769503	5	4.2E-04	NAHR(0.40),VNTR(0.10),NH(0.50),MEI(0.00)	
15	98809823	99309823	5	5.5E-03	NAHR(0.44),VNTR(0.00),NH(0.44),MEI(0.11)	15q26 overgrowth syndrome (chr15:97,175,493-100,338,915)
16	57172375	57704688	5	5.5E-03	NAHR(0.50),VNTR(0.10),NH(0.30),MEI(0.10)	
16	87335603	87925144	7	5.5E-03	NAHR(0.71),VNTR(0.21),NH(0.07),MEI(0.00)	
16	88097122	88597122	5	1.6E-03	NAHR(0.44),VNTR(0.22),NH(0.33),MEI(0.00)	
17	495756	1185675	7	3.8E-05	NAHR(0.92),VNTR(0.00),NH(0.08),MEI(0.00)	Miller-Dieker syndrome (chr17:1-2,492,179)
17	76270001	76949282	6	5.5E-03	NAHR(0.82),VNTR(0.09),NH(0.09),MEI(0.00)	
17	78165349	78665349	5	3.8E-05	NAHR(0.89),VNTR(0.00),NH(0.11),MEI(0.00)	
18	75158735	75910930	5	4.2E-04	NAHR(0.70),VNTR(0.10),NH(0.20),MEI(0.00)	
19	262971	840199	5	5.5E-03	NAHR(0.60),VNTR(0.10),NH(0.30),MEI(0.00)	
19	32549753	33049753	5	1.6E-03	NAHR(0.00),VNTR(0.78),NH(0.11),MEI(0.11)	
2	1079570	1853625	9	1.6E-03	NAHR(0.82),VNTR(0.12),NH(0.06),MEI(0.00)	
20	60476294	61256981	7	5.5E-03	NAHR(0.38),VNTR(0.23),NH(0.38),MEI(0.00)	
21	9795785	10332633	6	6.0E-06	NAHR(0.08),VNTR(0.67),NH(0.17),MEI(0.08)	
21	44977731	46668492	7	3.8E-05	NAHR(0.71),VNTR(0.00),NH(0.29),MEI(0.00)	
22	47023875	47535604	5	1.1E-04	NAHR(0.70),VNTR(0.10),NH(0.10),MEI(0.10)	
22	48008516	48595615	6	1.5E-05	NAHR(0.45),VNTR(0.09),NH(0.45),MEI(0.00)	
22	48855658	49355658*	5	1.1E-04	NAHR(0.56),VNTR(0.22),NH(0.22),MEI(0.00)	
4	1076778	1800747	5	1.6E-03	NAHR(0.44),VNTR(0.33),NH(0.22),MEI(0.00)	Wolf-Hirschhorn Syndrome (chr4:1-2,043,468)
4	7389047	8186412	6	4.2E-04	NAHR(0.58),VNTR(0.33),NH(0.08),MEI(0.00)	
4	15216519	15737675	5	5.5E-03	NAHR(0.10),VNTR(0.00),NH(0.90),MEI(0.00)	
4	190490833	191226269	7	5.5E-03	NAHR(0.50),VNTR(0.00),NH(0.33),MEI(0.17)	
5	1114210	1791351	6	1.1E-04	NAHR(0.83),VNTR(0.00),NH(0.17),MEI(0.00)	Cri du Chat Syndrome (5p deletion; chr5:1-11,776,854)
5	46013226	46513226	5	1.6E-03	NAHR(0.00),VNTR(0.67),NH(0.11),MEI(0.21)	
6	30948376	31494410	5	1.5E-05	NAHR(0.30),VNTR(0.00),NH(0.60),MEI(0.10)	
6	32421747	33701386	7	1.1E-04	NAHR(0.29),VNTR(0.00),NH(0.29),MEI(0.41)	
6	57347057	58031028	8	5.5E-03	NAHR(0.20),VNTR(0.00),NH(0.60),MEI(0.20)	
6	136623946	137131171	6	1.6E-03	NAHR(0.00),VNTR(0.00),NH(1.00),MEI(0.00)	
7	745210	1245210	5	3.8E-05	NAHR(0.56),VNTR(0.00),NH(0.22),MEI(0.22)	
7	154507924	155007924	5	5.5E-03	NAHR(0.44),VNTR(0.11),NH(0.33),MEI(0.11)	
7	157387155	158222236	7	7.2E-06	NAHR(0.77),VNTR(0.08),NH(0.08),MEI(0.08)	
8	950995	1540838	7	4.2E-04	NAHR(0.92),VNTR(0.00),NH(0.08),MEI(0.00)	
8	1888719	2609490	6	5.5E-03	NAHR(0.58),VNTR(0.08),NH(0.33),MEI(0.00)	
9	432598	998545	6	5.5E-03	NAHR(0.55),VNTR(0.00),NH(0.45),MEI(0.00)	
X	297290	1060682	5	3.8E-05	NAHR(0.50),VNTR(0.20),NH(0.10),MEI(0.20)	Leri-Weill dyschondrostosis - SHOX deletion (chrX:430,558-837,875)
X	1675344	1942198	5	1.5E-05	NAHR(0.50),VNTR(0.25),NH(0.25),MEI(0.00)	

**Near 22q13 deletion syndrome (chr2:49,392,382-49,534,710)*

Supplementary Table 15A. Accuracy of deletion calls with support from different SV discovery methods, low-coverage data

BC-RP	BI-PD	SD-RD	LN-SR	WU-RP	SI-RP	YL-RD	YL-RP	YL-SR	
0.111 (8)	0.091 (10)	NA (*)	0.333 (2)	0.167 (5)	0.143 (6)	NA (*)	0.200 (16)	0.500 (1)	**AE-RD**
	0.048 (118)	0.077 (12)	0.049 (97)	0.048 (119)	0.025 (116)	0.125 (7)	0.135 (45)	0.082 (45)	**BC-RP**
		0.045 (21)	0.049 (78)	0.051 (93)	0.035 (109)	0.111 (8)	0.136 (57)	0.065 (29)	**BI-PD**
			0.200 (4)	0.100 (9)	0.077 (12)	NA (*)	0.143 (30)	0.500 (1)	**SD-RD**
				0.034 (84)	0.013 (74)	0.333 (2)	0.130 (40)	0.095 (38)	**LN-SR**
					0.064 (88)	0.111 (8)	0.125 (42)	0.059 (32)	**WU-RP**
						0.250 (3)	0.169 (49)	0.029 (34)	**SI-RP**
							0.667 (2)	0.167 (5)	**YL-RD**
								0.208 (19)	**YL-RP**

(*) no SVs discovered by the respective methods

Fields in the half matrix above indicate the inferred FDR for SV calls with evidence from two methods. Numbers in parentheses indicate the number of successfully PCR-validated calls.

Supplementary Table 15B. Accuracy of deletion calls with support from different SV discovery methods, trio data

BG-AS	BC-RP	SD-RD	OX-AS	LN-SR	UW-RP	WU-RP	SI-RP	YL-RD	YL-RP	YL-SR	
0.059 (34)	0.048 (259)	0.038 (367)	0.333 (3)	0.050 (179)	0.052 (351)	0.058 (292)	0.068 (321)	0.037 (438)	0.156 (53)	0.104 (141)	**AB-RP**
	0.131 (59)	0.062 (31)	NA (*)	0.089 (41)	0.121 (58)	0.077 (60)	0.037 (53)	0.070 (40)	0.126 (15)	0.131 (53)	**BG-AS**
		0.034 (530)	0.116 (52)	0.047 (613)	0.043 (1082)	0.061 (801)	0.036 (918)	0.026 (708)	0.203 (87)	0.139 (519)	**BC-RP**
			0.214 (1899)	0.036 (339)	0.028 (748)	0.050 (479)	0.023 (637)	NA (*)	NA (*)	0.088 (1620)	**SD-RD**
				0.102 (40)	0.126 (51)	0.132 (44)	0.089 (23)	0.117 (46)	NA (*)	0.095 (46)	**OX-AS**
					0.059 (707)	0.056 (561)	0.043 (618)	0.032 (471)	0.165 (77)	0.121 (494)	**LN-SR**
						0.080 (1019)	0.072 (1097)	0.034 (972)	0.252 (133)	0.140 (606)	**UW-RP**
							0.100 (841)	0.035 (684)	0.248 (101)	0.148 (563)	**WU-RP**
								0.047 (822)	0.166 (121)	0.100 (508)	**SI-RP**
									NA (*)	0.283 (2854)	**YL-RD**
										NA (*)	**YL-RP**

(*) no SVs discovered by the respective methods

Fields in the half matrix above indicate the inferred FDR for SV calls with evidence from two methods. Numbers in parentheses indicate the number of successfully validated calls (array-CGH or PCR). We required at least 30 successful validation experiments and an estimated FDR$\leq$10% to consider a particular combination of methods in our analysis framework.

Supplementary Table 16. Effect of using subsets of deletion discovery methods along with the algorithm centric approach. The 5 top algorithms for the low-coverage set were (in order): BI-PD, BC-RP, SI-RP, LI-SR, and SD-RD. The top 5 algorithms for the trio set were (in order): BC-RP, AB-RP, SI-RP, LI-SR, and YL-RD.

	Number of callsets considered (ranked by FDR)				
Set	2	3	5	all	release set
low-coverage	8,930 (56%)	8,930 (56%)	9,117 (57%)	9,567 (60%)	15,947
trio	4,718 (42%)	4,825 (43%)	5,242 (46%)	7,336 (65%)	11,321
union	10,087 (46%)	10,140 (46%)	10,480 (48%)	12,652 (57%)	22,025

(*) Numbers in parenthesis represent percentage of release set recapitulated

Supplementary Table 17. Fraction of call set contributions ordered by set size

Set	Approach	Callset Origin	Total Set (%)
Low Coverage	PD	BI	44.3
	RD	SD	61.2
	SR	LI	78.6
	RP	SI	86.3
	RP	BC	91.1
	RP	WU	94.5
	RD	AE	97.3
	RP	YL	98.8
	RD	YL	99.9
	SR	YL	100
Trio	RP	BC	41.7
	RD	YL	55.7
	SR	YL	65.4
	RP	UW	74.1
	RD	SD	81.1
	RP	SI	87.5
	SR	LI	90.8
	RP	AB	94.1
	RP	WU	96.1
	AS	OX	97.8
	RD	UW	99.2
	RP	YL	99.9
	AS	BG	100

Supplementary Table 18. Summary of assembled breakpoints for deletion release set. The id column is linked to the ID field in the release VCF data set. The pos column indicates the reported breakpoint resolution coordinates for a particular deletion. The start and end columns list all assembled breakpoints associated for a particular deletion endpoint, stratified by approach type.

External File

Supplementary Table 19. Assessment of an enrichment of deletions intersecting with protein-coding sequences among highly differentiated deletions. Note that the three highly differentiated full gene deletions all fall into the CCL3L1 / CCL3L3 locus on chromosome 17.

Category		Highly differentiated*	All deletions	Chi squared p-value
complete gene deletion	yes	3	654	0.935731
	no	77	21371	
	ratios	0.04	0.03	
intersect gene	yes	32	9381	0.722708
	no	48	12644	
	ratios	0.67	0.74	
intersect cds	yes	2	1097	0.446479
	no	78	20928	
	ratios	0.03	0.05	
intersect utr	yes	1	315	0.736679
	no	79	21710	
	ratios	0.01	0.01	
intersect intron	yes	26	7319	0.984401
	no	54	14706	
	ratios	0.48	0.50	

*Fst >= 0.5 between two pairs of populations

Supplementary Table 20. Overlap of partial or whole genotyped, coding region deletions with OMIM Morbid Map. Coding regions were determined by comparing RefSeq genes with identified breakpoints (if available) or outer confidence intervals of genotyped deletions. These were then cross-linked with the OMIM Morbid Map database and the alternative allele and carrier frequencies were determined.

External File

References

[1] *Levy, S. et al. The diploid genome sequence of an individual human. PLoS Biol **5**, e254, doi:07-PLBI-RA-1258 [pii]*
10.1371/journal.pbio.0050254 (2007).

[2] *Wheeler, D. A. et al. The complete genome of an individual by massively parallel DNA sequencing. Nature **452**, 872-876, doi:nature06884 [pii]*
10.1038/nature06884 (2008).

[3] *Bentley, D. R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature **456**, 53-59, doi:nature07517 [pii]*
10.1038/nature07517 (2008).

[4] *McKernan, K. J. et al. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res **19**, 1527-1541, doi:gr.091868.109 [pii]*
10.1101/gr.091868.109 (2009).

[5] *Kim, J. I. et al. A highly annotated whole-genome sequence of a Korean individual. Nature **460**, 1011-1015, doi:nature08211 [pii]*
10.1038/nature08211 (2009).

[6] *Wang, J. et al. The diploid genome sequence of an Asian individual. Nature **456**, 60-65, doi:nature07484 [pii]*
10.1038/nature07484 (2008).

[7] *Conrad, D. F. et al. Origins and functional impact of copy number variation in the human genome. Nature **464**, 704-712, doi:nature08516 [pii]*
10.1038/nature08516 (2010).

[8] *Altshuler, D. M. et al. Integrating common and rare genetic variation in diverse human populations. Nature **467**, 52-58, doi:nature09298 [pii]*
10.1038/nature09298 (2010).

[9] *Redon, R. et al. Global variation in copy number in the human genome. Nature **444**, 444-454, doi:nature05329 [pii]*
10.1038/nature05329 (2006).

[10] *Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. B. CNVnator: an approach to characterize and genotype atypical CNVs using high-throughput sequencing coupled with population and family structure. submitted.*

[11] *Pang, A. W. et al. Towards a comprehensive structural variation map of an individual human genome. Genome Biol **11**, R52, doi:gb-2010-11-5-r52 [pii]*
10.1186/gb-2010-11-5-r52 (2010).

[12] *Abyzov, A. & Gerstein, M. B. Defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. submitted (2010).*

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

[14] *Chen, L. et al. TIGRA local targeted assembly of structural variants. submitted (2010).*

[15] *Yoon, S., Xuan, Z., Makarov, V., Ye, K. & Sebat, J. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res **19**, 1586-1592, doi:gr.092981.109 [pii]*
10.1101/gr.092981.109 (2009).

[16] *Korbel, J. O. et al. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol **10**, R23, doi:gb-2009-10-2-r23 [pii]*
10.1186/gb-2009-10-2-r23 (2009).

[17] *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:btp394 [pii]*

10.1093/bioinformatics/btp394 (2009).

[18] *Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. submitted.*

[19] *Alkan, C. et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet **41**, 1061-1067, doi:ng.437 [pii]*

10.1038/ng.437 (2009).

[20] *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:gr.088633.108 [pii]*

10.1101/gr.088633.108 (2009).

[21] *Li, R. et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res **20**, 265-272, doi:gr.097261.109 [pii]*

10.1101/gr.097261.109 (2010).

[22] *Hajirasouliha, I. et al. Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. Bioinformatics **26**, 1277-1283, doi:btq152 [pii]*

10.1093/bioinformatics/btq152 (2010).

[23] *Firth, H. V. et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am J Hum Genet **84**, 524-533, doi:S0002-9297(09)00107-4 [pii]*

10.1016/j.ajhg.2009.03.010 (2009).

The 1000 Genomes Project Consortium

Participants are arranged by project role, then by institution alphabetically, and finally alphabetically within institutions except for Principal Investigators and Project Leaders, as indicated.

Steering Committee: David L. Altshuler (Co-Chair)[2-4], Richard M. Durbin (Co-Chair)[1], Gonçalo R. Abecasis[5], David R. Bentley[6], Aravinda Chakravarti[7], Andrew G. Clark[8], Francis S. Collins[9], Francisco M. De La Vega[10], Peter Donnelly[11], Michael Egholm[12], Paul Flicek[13], Stacey B. Gabriel[2], Richard A. Gibbs[14], Bartha M. Knoppers[15], Eric S. Lander[2], Hans Lehrach[16], Elaine R. Mardis[17], Gil A. McVean[11,18], Debbie A. Nickerson[19], Leena Peltonen*, Alan J. Schafer[20], Stephen T. Sherry[21], Jun Wang[22,23], Richard K. Wilson[17]

Production Group: Baylor College of Medicine Richard A. Gibbs (Principal Investigator)[14], David Deiros[14], Mike Metzker[14], Donna Muzny[14], Jeff Reid[14], David Wheeler[14] **BGI-Shenzhen** Jun Wang (Principal Investigator) [22,23], Jingxiang Li[22], Min Jian[22], Guoqing Li[22], Ruiqiang Li[22,23], Huiqing Liang[22], Geng Tian[22], Bo Wang[22], Jian Wang[22], Wei Wang[22], Huanming Yang[22], Xiuqing Zhang[22], Huisong Zheng[22] **Broad Institute of MIT and Harvard** Eric S. Lander (Principal Investigator)[2], David L. Altshuler[2-4], Lauren Ambrogio[2], Toby Bloom[2], Kristian Cibulskis[2], Tim J. Fennell[2], Stacey B. Gabriel (Co-Chair)[2], David B. Jaffe[2], Erica Shefler[2], Carrie L. Sougnez[2] **Illumina** David R. Bentley (Principal Investigator)[6], Niall Gormley[6], Sean Humphray[6], Zoya Kingsbury[6], Paula Koko-Gonzales[6], Jennifer Stone[6] **Life Technologies** Kevin J. McKernan (Principal Investigator)[24], Gina L. Costa[24], Jeffry K. Ichikawa[24], Clarence C. Lee[24] **Max Planck Institute for Molecular Genetics** Ralf Sudbrak (Project Leader)[16], Hans Lehrach (Principal Investigator)[16], Tatiana A. Borodina[16], Andreas Dahl[25], Alexey N. Davydov[16], Peter Marquardt[16], Florian Mertes[16], Wilfiried Nietfeld[16], Philip Rosenstiel[26], Stefan Schreiber[26], Aleksey V. Soldatov[16], Bernd Timmermann[16], Marius Tolzmann[16] **Roche Applied Science** Michael Egholm (Principal Investigator)[12], Jason Affourtit[27], Dana Ashworth[27], Said Attiya[27], Melissa Bachorski[27], Eli Buglione[27], Adam Burke[27], Amanda Caprio[27], Christopher Celone[27], Shauna Clark[27], David Conners[27], Brian Desany[27], Lisa Gu[27], Lorri Guccione[27], Kalvin Kao[27], Andrew Kebbel[27], Jennifer Knowlton[27], Matthew Labrecque[27], Louise McDade[27], Craig Mealmaker[27], Melissa Minderman[27], Anne Nawrocki[27], Faheem Niazi[27], Kristen Pareja[27], Ravi Ramenani[27], David Riches[27], Wanmin Song[27], Cynthia Turcotte[27], Shally Wang[27] **Washington University in St. Louis** Elaine R. Mardis (Co-Chair) (Co-Principal Investigator) [17], Richard K. Wilson (Co-Principal Investigator)[17], David Dooling[17], Lucinda Fulton[17], Robert Fulton[17], George Weinstock[17] **Wellcome Trust Sanger Institute** Richard M. Durbin (Principal Investigator)[1], John Burton[1], David M. Carter[1], Carol Churcher[1], Alison Coffey[1], Anthony Cox[1], Aarno Palotie[1,28], Michael Quail[1], Tom Skelly[1], James Stalker[1], Harold P. Swerdlow[1], Daniel Turner[1]

Analysis Group: Agilent Technologies Anniek De Witte[29], Shane Giles[29] **Baylor College of Medicine** Richard A. Gibbs (Principal Investigator)[14], David Wheeler[14], Matthew Bainbridge[14], Danny Challis[14], Aniko Sabo[14], Fuli Yu[14], Jin Yu[14] **BGI-Shenzhen** Jun Wang (Principal Investigator)[22,23], Xiaodong Fang[22], Xiaosen Guo[22], Ruiqiang Li[22,23], Yingrui Li[22], Ruibang Luo[22], Shuaishuai Tai[22], Honglong Wu[22], Hancheng Zheng[22], Xiaole Zheng[22], Yan Zhou[22], Guoqing Li[22], Jian Wang[22], Huanming Yang[22] **Boston College** Gabor T. Marth (Principal Investigator)[30], Erik P. Garrison[30], Weichun Huang[31], Amit Indap[30], Deniz Kural[30], Wan-Ping Lee[30], Wen Fung Leong[30], Aaron R. Quinlan[32], Chip Stewart[30], Michael P. Stromberg[33], Alistair N. Ward[30], Jiantao Wu[30] **Brigham and Women's Hospital** Charles Lee (Principal Investigator)[34], Ryan E. Mills[34], Xinghua Shi[34] **Broad Institute of MIT and Harvard** Mark J. Daly (Principal Investigator)[2], Mark A. DePristo (Project Leader)[2], David L. Altshuler[2-4], Aaron D. Ball[2], Eric Banks[2], Toby Bloom[2], Brian L. Browning[35], Kristian Cibulskis[2], Tim J. Fennell[2], Kiran V. Garimella[2], Sharon R. Grossman[2,36], Robert E. Handsaker[2], Matt Hanna[2], Chris Hartl[2], David B. Jaffe[2], Andrew M. Kernytsky[2], Joshua M. Korn[2], Heng Li[2], Jared R. Maguire[2], Steven A. McCarroll[2,4], Aaron McKenna[2], James C. Nemesh[2], Anthony A. Philippakis[2], Ryan E. Poplin[2], Alkes Price[37], Manuel A. Rivas[2], Pardis C. Sabeti[2,36], Stephen F. Schaffner[2], Erica Shefler[2], Ilya A. Shlyakhter[2,36] **Cardiff University, The Human Gene Mutation Database** David N. Cooper (Principal Investigator)[38], Edward V. Ball[38], Matthew Mort[38], Andrew D. Phillips[38], Peter D. Stenson[38]

Cold Spring Harbor Laboratory Jonathan Sebat (Principal Investigator)[39], Vladimir Makarov[40], Kenny Ye[41], Seungtai C. Yoon[42] **Cornell and Stanford Universities** Carlos D. Bustamante (Co-Principal Investigator)[43], Andrew G. Clark (Co-Principal Investigator)[8], Adam Boyko[43], Jeremiah Degenhardt[8], Simon Gravel[43], Ryan N. Gutenkunst[44], Mark Kaganovich[43], Alon Keinan[8], Phil Lacroute[43], Xin Ma[8], Andy Reynolds[8] **European Bioinformatics Institute** Laura Clarke (Project Leader)[13], Paul Flicek (Co-Chair, DCC) (Principal Investigator)[13], Fiona Cunningham[13], Javier Herrero[13], Stephen Keenen[13], Eugene Kulesha[13], Rasko Leinonen[13], William M. McLaren[13], Rajesh Radhakrishnan[13], Richard E. Smith[13], Vadim Zalunin[13], Xiangqun Zheng-Bradley[13] **European Molecular Biology Laboratory** Jan O. Korbel (Principal Investigator)[45], Adrian M. Stütz[45] **Illumina** Sean Humphray (Project Leader)[6], Markus Bauer[6], R. Keira Cheetham[6], Tony Cox[6], Michael Eberle[6], Terena James[6], Scott Kahn[6], Lisa Murray[6] **Johns Hopkins University** Aravinda Chakravarti[7] **Leiden University Medical Center** Kai Ye[46] **Life Technologies** Francisco M. De La Vega (Principal Investigator)[10], Yutao Fu[24], Fiona C.L. Hyland[10], Jonathan M. Manning[24], Stephen F. McLaughlin[24], Heather E. Peckham[24], Onur Sakarya[10], Yongming A. Sun[10], Eric F. Tsung[24] **Louisiana State University** Mark A. Batzer (Principal Investigator)[47], Miriam K. Konkel[47], Jerilyn A. Walker[47] **Max Planck Institute for Molecular Genetics** Ralf Sudbrak (Project Leader)[16], Marcus W. Albrecht[16], Vyacheslav S. Amstislavskiy[16], Ralf Herwig[16], Dimitri V. Parkhomchuk[16] **US National Institutes of Health** Stephen T. Sherry (Co-Chair, DCC) (Principal Investigator)[21], Richa Agarwala[21], Hoda M. Khouri[21], Aleksandr O. Morgulis[21], Justin E. Paschall[21], Lon D. Phan[21], Kirill E. Rotmistrovsky[21], Robert D. Sanders[21], Martin F. Shumway[21], Chunlin Xiao[21] **Oxford University** Gil A. McVean (Co-Chair) (Co-Chair, Population Genetics) (Principal Investigator)[11,18], Adam Auton[11], Zamin Iqbal[11], Gerton Lunter[11], Jonathan L. Marchini[11,18], Loukas Moutsianas[18], Simon Myers[11,18], Afidalina Tumian[18] **Roche Applied Science** Brian Desany (Project Leader)[27], James Knight[27], Roger Winer[27] **The Translational Genomics Research Institute** David W. Craig (Principal Investigator)[48], Steve M. Beckstrom-Sternberg[48], Alexis Christoforides[48], Ahmet A. Kurdoglu[48], John V. Pearson[48], Shripad A. Sinari[48], Waibhav D. Tembe[48] **University of California, Santa Cruz** David Haussler (Principal Investigator)[49], Angie S. Hinrichs[49], Sol J. Katzman[49], Andrew Kern[49], Robert M. Kuhn[49] **University of Chicago** Molly Przeworski (Co-Chair, Population Genetics) (Principal Investigator)[50], Ryan D. Hernandez[51], Bryan Howie[52], Joanna L. Kelley[52], S. Cord Melton[52] **University of Michigan** Gonçalo R. Abecasis (Co-Chair) (Principal Investigator)[5], Yun Li (Project Leader)[5], Paul Anderson[5], Tom Blackwell[5], Wei Chen[5], William O. Cookson[53], Jun Ding[5], Hyun Min Kang[5], Mark Lathrop[54], Liming Liang[55], Miriam F. Moffatt[53], Paul Scheet[56], Carlo Sidore[5], Matthew Snyder[5], Xiaowei Zhan[5], Sebastian Zöllner[5] **University of Montreal** Philip Awadalla (Principal Investigator)[57], Ferran Casals[58], Youssef Idaghdour[58], John Keebler[58], Eric A. Stone[58], Martine Zilversmit[58] **University of Utah** Lynn Jorde (Principal Investigator)[59], Jinchuan Xing[59] **University of Washington** Evan E. Eichler (Principal Investigator)[60], Gozde Aksay[19], Can Alkan[60], Iman Hajirasouliha[61], Fereydoun Hormozdiari[61], Jeffrey M. Kidd[19,43], S. Cenk Sahinalp[61], Peter H. Sudmant[19] **Washington University in St. Louis** Elaine R. Mardis (Co-Principal Investigator)[17], Ken Chen[17], Asif Chinwalla[17], Li Ding[17], Daniel C. Koboldt[17], Mike D. McLellan[17], David Dooling[17], George Weinstock[17], John W. Wallis[17], Michael C. Wendl[17], Qunyuan Zhang[17] **Wellcome Trust Sanger Institute** Richard M. Durbin (Principal Investigator)[1], Cornelis A. Albers[62], Qasim Ayub[1], Senduran Balasubramaniam[1], Jeffrey C. Barrett[1], David M. Carter[1], Yuan Chen[1], Donald F. Conrad[1], Petr Danecek[1], Emmanouil T. Dermitzakis[63], Min Hu[1], Ni Huang[1], Matt E. Hurles[1], Hanjun Jin[64], Luke Jostins[1], Thomas M. Keane[1], Si Quang Le[1], Sarah Lindsay[1], Quan Long[1], Daniel G. MacArthur[1], Stephen B. Montgomery[63], Leopold Parts[1], James Stalker[1], Chris Tyler-Smith[1], Klaudia Walter[1] Yujun Zhang[1] **Yale and Stanford Universities** Mark B. Gerstein (Co-Principal Investigator)[65,66], Michael Snyder (Co-Principal Investigator)[43], Alexej Abyzov[65], Suganthi Balasubramanian[67], Robert Bjornson[66], Jiang Du[66], Fabian Grubert[43], Lukas Habegger[65], Rajini Haraksingh[65], Justin Jee[65], Ekta Khurana[67], Hugo Y.K. Lam[43], Jing Leng[65], Xinmeng Jasmine Mu[65], Alexander E. Urban[43,68], Zhengdong Zhang[67]

Structural Variation Group: BGI-Shenzhen Yingrui Li[22], Ruibang Luo[22] **Boston College** Gabor T. Marth (Principal Investigator)[30], Erik P. Garrison[30], Deniz Kural[30], Aaron R. Quinlan[32], Chip Stewart[30], Michael P.

Stromberg[33], Alistair N. Ward[30], Jiantao Wu[30] **Brigham and Women's Hospital** Charles Lee (Co-Chair) (Principal Investigator)[34], Ryan E. Mills[34], Xinghua Shi[34] **Broad Institute of MIT and Harvard** Steven A. McCarroll (Project Leader)[2,4], Eric Banks[2], Mark A. DePristo[2], Robert E. Handsaker[2], Chris Hartl[2], Joshua M. Korn[2], Heng Li[2], James C. Nemesh[2] **Cold Spring Harbor Laboratory** Jonathan Sebat (Principal Investigator)[39], Vladimir Makarov[40], Kenny Ye[41], Seungtai C. Yoon[42] **Cornell and Stanford Universities** Jeremiah Degenhardt[8], Mark Kaganovich[43] **European Bioinformatics Institute** Laura Clarke (Project Leader)[13], Richard E. Smith[13], Xiangqun Zheng-Bradley[13] **European Molecular Biology Laboratory** Jan O. Korbel[45] **Illumina** Sean Humphray (Project Leader)[6], R. Keira Cheetham[6], Michael Eberle[6], Scott Kahn[6], Lisa Murray[6] **Leiden University Medical Center** Kai Ye[46] **Life Technologies** Francisco M. De La Vega (Principal Invesigator)[10], Yutao Fu[24], Heather E. Peckham[24], Yongming A. Sun[10] **Louisiana State University** Mark A. Batzer (Principal Investigator)[47], Miriam K. Konkel[47], Jerilyn A. Walker[47] **US National Institutes of Health** Chunlin Xiao[21] **Oxford University** Zamin Iqbal[11] **Roche Applied Science** Brian Desany[27] **University of Michigan** Tom Blackwell (Project Leader)[5], Matthew Snyder[5] **University of Utah** Jinchuan Xing[59] **University of Washington** Evan E. Eichler (Co-Chair) (Principal Investigator)[60], Gozde Aksay[19], Can Alkan[60], Iman Hajirasouliha[61], Fereydoun Hormozdiari[61], Jeffrey M. Kidd[19,43] **Washington University in St. Louis** Ken Chen[17], Asif Chinwalla[17], Li Ding[17], Mike D. McLellan[17], John W. Wallis[17] **Wellcome Trust Sanger Institute** Matt E. Hurles[1] (Co-Chair) (Principal Investigator), Donald F. Conrad[1], Klaudia Walter[1], Yujun Zhang[1] **Yale and Stanford Universities** Mark B. Gerstein (Co-Principal Investigator)[65,66], Michael Snyder (Co-Principal Investigator)[43], Alexej Abyzov[65], Jiang Du[66], Fabian Grubert[43], Rajini Haraksingh[65], Justin Jee[65], Ekta Khurana[67], Hugo Y.K. Lam[43], Jing Leng[65], Xinmeng Jasmine Mu[65], Alexander E. Urban[43,68], Zhengdong Zhang[67]

Exon Pilot Group: Baylor College of Medicine Richard A. Gibbs (Co-Chair) (Principal Investigator)[14], Matthew Bainbridge[14], Danny Challis[14], Cristian Coafra[14], Huyen Dinh[14], Christie Kovar[14], Sandy Lee[14], Donna Muzny[14], Lynne Nazareth[14], Jeff Reid[14], Aniko Sabo[14], Fuli Yu[14], Jin Yu[14] **Boston College** Gabor T. Marth (Co-Chair) (Principal Investigator)[30], Erik P. Garrison[30], Amit Indap[30], Wen Fung Leong[30], Aaron R. Quinlan[32], Chip Stewart[30], Alistair N. Ward[30], Jiantao Wu[30] **Broad Institute of MIT and Harvard** Kristian Cibulskis[2], Tim J. Fennell[2], Stacey B. Gabriel[2], Kiran V. Garimella[2], Chris Hartl[2], Erica Shefler[2], Carrie L. Sougnez[2], Jane Wilkinson[2] **Cornell and Stanford Universities** Andrew G. Clark (Co-Principal Investigator)[8], Simon Gravel[43], Fabian Grubert[43] **European Bioinformatics Institute** Laura Clarke (Project Leader)[13], Paul Flicek (Principal Investigator)[13], Richard E. Smith[13], Xiangqun Zheng-Bradley[13] **US National Institutes of Health** Stephen T. Sherry (Principal Investigator)[21], Hoda M. Khouri[21], Justin E. Paschall[21], Martin F. Shumway[21], Chunlin Xiao[21] **Oxford University** Gil A. McVean[11,18] **University of California, Santa Cruz** Sol J. Katzman[49] **University of Michigan** Gonçalo R. Abecasis (Co-Chair) (Principal Investigator)[5], Tom Blackwell[5] **Washington University in St. Louis** Elaine R. Mardis (Principal Investigator)[17], David Dooling[17], Lucinda Fulton[17], Robert Fulton[17], Daniel C. Koboldt[17] **Wellcome Trust Sanger Institute** Richard M. Durbin (Principal Investigator)[1], Senduran Balasubramaniam[1], Allison Coffey[1], Thomas M. Keane[1], Daniel G. MacArthur[1], Aarno Palotie[1,28], Carol Scott[1], James Stalker[1], Chris Tyler-Smith[1] **Yale University** Mark B. Gerstein (Principal Investigator)[65,66], Suganthi Balasubramanian[67]

Samples and ELSI Group: Aravinda Chakravarti (Co-Chair)[7], Bartha M. Knoppers (Co-Chair)[15], Leena Peltonen (Co-Chair)*, Gonçalo R. Abecasis[5], Carlos D. Bustamante[43], Neda Gharani[69], Richard A. Gibbs[14], Lynn Jorde[59], Jane S. Kaye[70], Alastair Kent[71], Taosha Li[22], Amy L. McGuire[72], Gil A. McVean[11,18], Pilar N. Ossorio[73], Charles N. Rotimi[74], Yeyang Su[22], Lorraine H. Toji[69], Chris Tyler-Smith[1]

Scientific Management: Lisa D. Brooks[75], Adam L. Felsenfeld[75], Jean E. McEwen[75], Assya Abdallah[76], Christopher R. Juenger[77], Nicholas C. Clemm[75], Francis S. Collins[9], Audrey Duncanson[20], Eric D. Green[78], Mark S. Guyer[75], Jane L. Peterson[75], Alan J. Schafer[20]

1 Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA,UK.

2 The Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA.

3 Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts 02114, USA.

4 Dept of Genetics, Harvard Medical School, Cambridge, Massachusetts 02115 , USA.

5 Center for Statistical Genetics and Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA.

6 Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, Nr Saffron Walden, Essex CB10 1XL, UK.

7 McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA.

8 Center for Comparative and Population Genomics, Cornell University, Ithaca, New York 14850, USA.

9 US National Institutes of Health, 1 Center Drive, Bethesda, Maryland 20892, USA.

10 Life Technologies, Foster City, California 94404, USA.

11 Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK.

12 Pall Corporation, 25 Harbor Park Drive, Port Washington, New York 11050 USA.

13 European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SD, UK.

14 Human Genome Sequencing Center, Baylor College of Medicine, 1 Baylor Plaza, Houston, Texas 77030, USA.

15 Centre of Genomics and Policy, McGill University, Montréal, Québec H3A 1A4, Canada.

16 Max Planck Institute for Molecular Genetics, D-14195 Berlin-Dahlem, Germany.

17 The Genome Center, Washington University School of Medicine, St Louis, Missouri 63108, USA.

18 Dept of Statistics, University of Oxford, Oxford OX1 3TG, UK.

19 Dept of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA.

20 Wellcome Trust, Gibbs Building, 215 Euston Road, London NW1 2BE, UK.

21 US National Institutes of Health, National Center for Biotechnology Information, 45 Center Drive, Bethesda, Maryland 20892, USA.

22 BGI-Shenzhen, Shenzhen 518083, China.

23 Dept of Biology, University of Copenhagen, Denmark.

24 Life Technologies, Beverly, Massachussetts 01915, USA.

25 Deep Sequencing Group, Biotechnology Center TU Dresden, Tatzberg 47/49, 01307, Dresden, Germany.

26 Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, Kiel, Germany.

27 Roche Applied Science, 20 Commercial Street, Branford, Connecticut 06405, USA.

28 Department of Medical Genetics, Institute of Molecular Medicine (FIMM) of the University of Helsinki and Helsinki University Hospital, Helsinki, Finland.

29 Agilent Technologies Inc., Santa Clara, California 95051, USA.

30 Dept of Biology, Boston College, Chestnut Hill, Massachusetts 02467, USA.

31 US National Institutes of Health, National Institute of Environmental Health Sciences, 111 T W, Alexander Drive, Research Triangle Park, North Carolina 27709, USA.

32 Dept of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, Charlottesville,

Virginia 22908, USA.

33 Illumina, San Diego, California 92121, USA.

34 Dept of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA.

35 Dept of Medicine, Division of Medical Genetics, University of Washington, Seattle, Washington 98195, USA.

36 Center for Systems Biology, Dept Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachussetts 02138, USA.

37 Dept of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, USA.

38 Institute of Medical Genetics, Cardiff University, Heath Park, Cardiff CF14 4XN, UK.

39 Depts of Psychiatry and Cellular and Molecular Medicine, University of California San Diego, 9500 Gilman Dr, La Jolla, California 92093, USA.

40 Seaver Autism Center and Department of Psychiatry, Mount Sinai School of Medicine, New York, New York 10029, USA.

41 Dept of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, New York 10461, USA.

42 Dept of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, New York 10029, USA.

43 Dept of Genetics, Stanford University, Stanford, California 94305, USA.

44 Dept of Molecular and Cellular Biology, University of Arizona, Tucson, Arizona 85721, USA.

45 European Molecular Biology Laboratory, Genome Biology Research Unit, Meyerhofstr. 1, Heidelberg, Germany.

46 Molecular Epidemiology Section, Medical Statistics and Bioinformatics, Leiden University Medical Center, 2333 ZA, The Netherlands.

47 Dept of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana 70803, USA.

48 The Translational Genomics Research Institute, 445 N Fifth Street, Phoenix, Arizona 85004, USA.

49 Center for Biomolecular Science and Engineering, University of California Santa Cruz, Santa Cruz, California 95064, USA.

50 Dept of Human Genetics and Howard Hughes Medical Institute, University of Chicago, Chicago, Illinois 60637, USA.

51 Dept of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California 94158, USA.

52 Dept of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA.

53 National Heart and Lung Institute, Imperial College London, London SW7 2, UK.

54 Centre Nationale de Génotypage, Evry, France.

55 Depts of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, USA.

56 Dept of Epidemiology, University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA.

57 Dept of Pediatrics, Faculty of Medicine, University of Montréal, Ste. Justine Hospital Research Centre, Montréal, Québec H3T 1C5, Canada.

58 Dept of Medicine, Centre Hospitalier de l'Université de Montréal Research Center, Université de Montréal,

Montréal, Québec H2L 2W5, Canada.

59 Eccles Institute of Human Genetics, University of Utah School of Medicine, Salt Lake City, Utah 84112, USA.

60 Dept of Genome Sciences, University of Washington School of Medicine and Howard Hughes Medical Institute, Seattle, Washington 98195, USA.

61 Dept of Computer Science, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada.

62 Dept of Haematology, University of Cambridge and National Health Service Blood and Transplant, Cambridge CB2 1TN, UK.

63 Dept of Genetic Medicine and Development, University of Geneva Medical School, Geneva, 1211 Switzerland.

64 Center for Genome Science, Korea National Institute of Health, 194, Tongil-Lo, Eunpyung-Gu, Seoul, 122-701, Korea.

65 Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA.

66 Dept of Computer Science, Yale University, New Haven, Connecticut 06520, USA.

67 Dept of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA.

68 Dept of Psychiatry and Behavioral Studies, Stanford University, Stanford, California 94305, USA.

69 Coriell Institute, 403 Haddon Avenue, Camden, New Jersey 08103, USA.

70 Centre for Health, Law and Emerging Technologies, University of Oxford, Oxford OX3 7LF, UK.

71 Genetic Alliance, 436 Essex Road, London, N1 3QP, UK.

72 Center for Medical Ethics and Health Policy, Baylor College of Medicine, 1 Baylor Plaza, Houston, Texas 77030, USA.

73 Dept of Medical History and Bioethics, University of Wisconsin--Madison, Madison, Wisconsin 53706, USA.

74 US National Institutes of Health, Center for Research on Genomics and Global Health, 12 South Drive, Bethesda, Maryland 20892, USA.

75 US National Institutes of Health, National Human Genome Research Institute, 5635 Fishers Lane, Bethesda, Maryland 20892, USA.

76 The George Washington University School of Medicine and Health Sciences, Washington, DC 20037, USA.

77 US Food and Drug Administration, 11400 Rockville Pike, Rockville, Maryland 20857, USA.

78 US National Institutes of Health, National Human Genome Research Institute, 31 Center Drive, Bethesda, Maryland 20892, USA.

* Deceased.