ARTICLE

doi:10.1038/nature09708

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[13,27], 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†

Genomic structural variants (SVs) are abundant in humans, differing from other forms of variation in extent, origin and functional impact. Despite progress in SV characterization, the nucleotide resolution architecture of most SVs remains unknown. We constructed a map of unbalanced SVs (that is, copy number variants) based on whole genome DNA sequencing data from 185 human genomes, integrating evidence from complementary SV discovery approaches with extensive experimental validations. Our map encompassed 22,025 deletions and 6,000 additional SVs, including insertions and tandem duplications. Most SVs (53%) were mapped to nucleotide resolution, which facilitated analysing their origin and functional impact. We examined numerous whole and partial gene deletions with a genotyping approach and observed a depletion of gene disruptions amongst high frequency deletions. Furthermore, we observed differences in the size spectra of SVs originating from distinct formation mechanisms, and constructed a map of SV hotspots formed by common mechanisms. Our analytical framework and SV map serves as a resource for sequencing-based association studies.

Introduction

Unbalanced structural variants (SVs), or copy number variants (CNVs), involving large-scale deletions, duplications and insertions form one of the least well studied classes of genetic variation. The fraction of the genome affected by SVs is comparatively larger than that accounted for by single nucleotide polymorphisms[1] (SNPs), implying significant consequences of SVs on phenotypic variation. SVs have already been associated with diverse diseases, including autism[2,3], schizophrenia[4,5] and Crohn's disease[6,7]. Furthermore, locus-specific studies suggest that diverse mechanisms may form SVs *de novo*, with some mechanisms involving complex rearrangements resulting in multiple chromosomal breakpoints[8,9].

Initial microarray-based SV surveys focused on large gains and losses[10–12], with recent advances in array technology widening the accessible size spectrum towards smaller SVs[1,13]. Microarray-based surveys commonly mapped SVs to approximate genomic locations. However, a detailed SV characterization, including analyses of SV origin and impact, requires knowledge of precise SV sequences. Advances in sequencing technology have enabled applying sequence-based approaches for mapping SVs at a fine scale[14–21]. These approaches include: (1) paired-end mapping (or read pair 'RP' analysis) based on sequencing and analysis of abnormally mapping pairs of clone ends[14,22–24] or high-throughput sequencing fragments[15,17,18]; (2) read-depth ('RD') analysis, which detects SVs by analysing the read depth-of-coverage[16,21,25–27]; (3) split-read ('SR') analysis, which evaluates gapped sequence alignments for SV detection[28,29]; and (4) sequence assembly ('AS'), which enables the fine-scale discovery of SVs, including novel (non-reference) sequence insertions[30–32]. Sequence-based SV discovery approaches have previously been applied to a limited (<20) number of genomes, leaving the fine-scale architecture of most common SVs unknown.

Sequence data generated by the 1000 Genomes Project (1000GP) provide an unprecedented opportunity to generate a comprehensive SV map. The 1000GP recently generated 4.1 terabases of raw sequence

[1]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA. [2]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK. [3]Department of Biology, Boston College, Boston, Massachusetts, USA. [4]Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]The Genome Center at Washington University, St. Louis, Missouri, USA. [6]Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington, USA. [7]Howard Hughes Medical Institute, University of Washington, Seattle, Washington, USA. [8]Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, USA. [9]Seaver Autism Center and Department of Psychiatry, Mount Sinai School of Medicine, New York, New York, USA. [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, Saffron Walden CB10 1XL, UK. [12]Life Technologies, Beverly, Massachusetts, USA. [13]Department of Genetics, Stanford University, Stanford, California, USA. [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, California, USA. [16]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. [17]Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana, USA. [18]Molecular Biophysics and Biochemistry Department, Yale University, New Haven, Connecticut, USA. [19]BGI-Shenzhen, Shenzhen 518083, China. [20]Albert Einstein College of Medicine, Bronx, New York, USA. [21]Genome Biology Research Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [22]Department of Genetics, Washington University, St Louis, Missouri, USA. [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, Connecticut, USA. [26]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [27]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, California, USA.

*These authors contributed equally to this work.

†Lists of participants and affiliations are shown in Supplementary Information.

in two pilot projects targeting whole human genomes[33] (Supplementary Table 1). These studies comprise a population-scale project, termed 'low-coverage project', in which 179 unrelated individuals were sequenced with an average coverage of 3.6×, including 59 Yoruba individuals from Nigeria (YRI), 60 individuals of European ancestry from Utah (CEU), 30 of Han ancestry from Beijing (CHB), and 30 of Japanese ancestry from Tokyo (JPT; the latter two were jointly analysed as JPT+CHB). In addition, a high-coverage project, termed the 'trio project', was carried out, with individuals of a CEU and a YRI parent-offspring trio sequenced to 42× coverage on average.

We report here the results of analyses undertaken by the Structural Variation Analysis Group of the 1000GP. The group's objectives were to discover, assemble, genotype and validate SVs of 50 base pairs (bp) and larger in size, and to assess and compare different sequence-based SV detection approaches. The focus of the group was initially on deletions, a variant class often associated with disease[9], for which rich control data sets and diverse ascertainment approaches exist[1,13,22,28]. Less focus was placed on insertions and duplications[34] and none on balanced SV forms (such as inversions). Specifically, we applied nineteen methods to generate an SV discovery set. We further generated reference genotypes for most deletions, assessed the SVs' functional impact and stratified SV formation mechanism with respect to variant size and genomic context.

Assessment of SV discovery methods

We incorporated the SV discovery methods into a pipeline (Fig. 1a, b), with the goal of ascertaining different SV types and assessing each method for its ability to discover SVs. The methods detected SVs by analysing RD, RP, SR and AS features, or by combining RP and RD features (abbreviated as 'PD'). Altogether we generated 36 SV call-sets by applying the methods on trio and low-coverage whole genome sequence data, and by identifying SVs as genomic differences relative to a human reference, corresponding to the reference genome, or to a set of individuals (that is, population reference; Supplementary Table 2). We initially identified SVs as deletions, tandem duplications, novel sequence insertions and mobile element insertions (MEIs) relative to the human reference. Subsequent comparative analyses involving primate genomes enabled us to classify SVs as deletions, duplications, or insertions relative to inferred ancestral genomic loci, reflecting mechanisms of SV formation (see below). DNA reads analysed by SV discovery methods were initially mapped to the human reference genome using a variety of alignment algorithms. Most of these algorithms mapped each read to a single genomic position, although one algorithm (mrFAST[16]) also considered alternative mapping positions for reads aligning to repetitive regions (see Supplementary Tables 2–4 for method-specific parameters and full SV call-sets). We filtered each call-set by excluding SVs <50 bp, which are reported elsewhere[33]. Many SVs showed support from distinct SV discovery methods, as exemplified by a common deletion, previously associated with body-mass index[35] (BMI), that we identified with RP, RD and SR methods (Fig. 1c). Nonetheless, we observed notable differences between methods (Fig. 2a–c) in terms of genomic regions ascertained (Supplementary Fig. 1), accessible SV size-range (Fig. 2a), and breakpoint precision (Fig. 2c, Supplementary Fig. 2).

To estimate call-set specificity, we carried out extensive validations (Methods), including PCRs for over 3,000 candidate loci and microarray data analyses for 50,000 candidate loci (Supplementary Tables 3, 4 and Supplementary Fig. 3). We combined PCR and array-based analysis results to estimate false discovery rates (FDRs), and found that eight call-sets (three deletion, one tandem duplication and four insertion call-sets) met the pre-specified specificity threshold[33] (FDR ≤ 10%), whereas the other call-sets yielded lower specificity (FDRs of 13–89%).

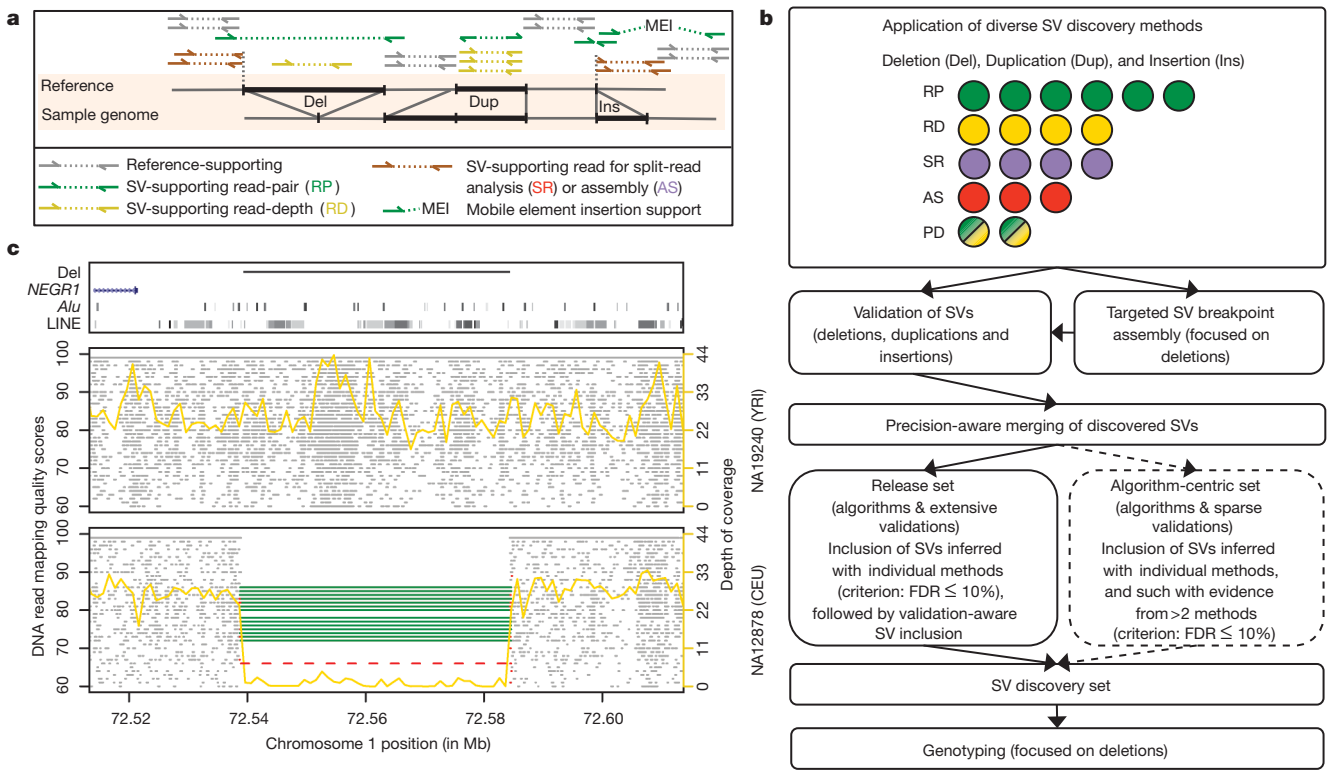

Figure 1 | SV discovery and genotyping in population scale sequence data. **a**, Schematic depicting the different modes (that is, approaches) of sequence-based SV detection we used. The RP approach assesses the orientation and spacing of the mapped reads of paired-end sequences[14,15] (reads are denoted by arrows); the RD approach evaluates the read depth-of-coverage[25,26]; the SR approach maps the boundaries (breakpoints) of SVs by sequence alignment[28,29]; the AS approach assembles SVs[30–32]. **b**, Integrated pipeline for SV discovery, validation and genotyping. Coloured circles represent individual SV discovery methods (listed in Supplementary Table 2), with modes indicated by a colour scheme: green, RP; yellow, RD; purple, SR; red, AS; green and yellow, methods evaluating RP and RD (abbreviated as 'PD'). **c**, Example of a deletion, previously associated with BMI[35], identified independently with RP (green), RD (yellow) and SR (red) methods. Grey dots indicate position and mapping quality for individual sequence reads. Targeted assembly confirmed the breakpoints detected by SR.

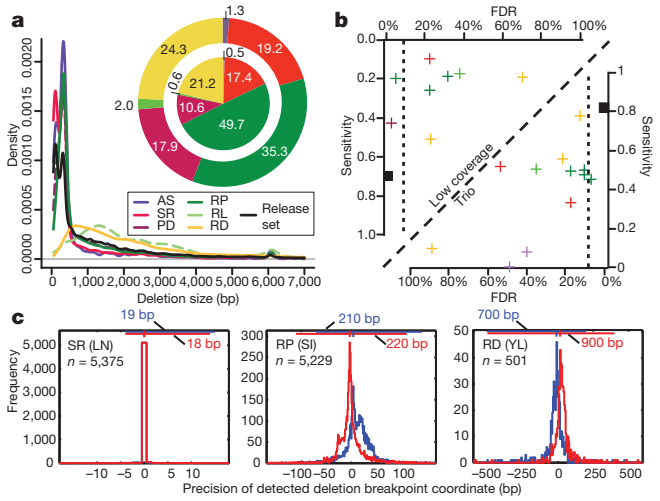

Figure 2 | Comparative assessment of deletion discovery methods. **a**, Deletion size-range ascertained by different modes of SV discovery. Three groups are visible, with AS and SR, PD and RP, as well as RD and 'RL' (RP analysis involving relatively long range (≥1 kb) insert size libraries, resulting in a different deletion detection size range compared to the predominantly used <500 bp insert size libraries), respectively, ascertaining similar size-ranges. Pie charts display the contribution (%) of different SV discovery modes to the release set. Outer pie is based on the number of SV calls; inner pie is based on the total number of variable nucleotides. Of note, not all approaches were applied across all individuals (see Supplementary Table 2). **b**, Sensitivity and FDR estimates for individual deletion discovery methods based on gold standard sets for individuals sequenced at high (NA12878) and low-coverage (NA12156), respectively. All depicted estimates are summarized in Supplementary Tables 3, 4 and 6. Vertical dotted lines correspond to the specificity threshold (FDR ≤ 10%). **c**, Breakpoint mapping resolution of three deletion discovery methods (the respective method names are in Supplementary Table 2). The blue and red histograms are the breakpoint residuals for predicted deletion start and end coordinates, respectively, relative to assembled coordinates (here assessed in low-coverage data). The horizontal lines at the top of each plot mark the 98% confidence intervals (labelled for each panel), with vertical notches indicating the positions of the most probable breakpoint (the distribution mode).

We assessed the sensitivity of deletion discovery methods further by collating data from four earlier surveys[1,13,22,28] into a gold standard (Methods, Supplementary Tables 5, 6 and Supplementary Fig. 4a), and specifically assessing the detection sensitivity for an individual sequenced at high-coverage (NA12878) as well as for an individual sequenced at low-coverage (NA12156). Unsurprisingly, given the typical trade-off between sensitivity and specificity, in the trios the highest sensitivities were achieved by RD and RP methods with FDR > 10% (Fig. 2b). By comparison, in the low-coverage data, the individual method with the greatest accuracy (FDR = 3.7%) was the second most sensitive based on our gold standard (Fig. 2b), and the most sensitive when expanding the gold standard to a larger set of individuals (Supplementary Fig. 4b). This method, Genome STRiP (to be described elsewhere; Handsaker, R. E., Korn, J. M., Nemesh, J. and McCarroll, S. A., unpublished results), integrated both RP and RD features (PD), implying that considering different evidence types can improve SV discovery.

Construction of our SV discovery set

To construct our SV discovery set ('release set'), we joined calls from different discovery methods corresponding to the same SV with a merging approach that was aware of each call-set's precision in SV breakpoint detection (Supplementary Fig. 5 and Methods). Most SVs in the release set (61%) were contributed by individual methods meeting the pre-defined specificity threshold (FDR ≤ 10%). The remaining 39% of calls were contributed by lower specificity methods following experimental validation. Altogether, the release set comprised 22,025 deletions, 501 tandem duplications, 5,371 MEIs and 128 non-reference insertions (Table 1, Supplementary Table 7). With our gold standard we estimated an overall sensitivity of deletion discovery of 82% in the trios, and 69% in low-coverage sequence (Fig. 2b) using a 1-bp overlap criterion. When instead applying a stringent 50% reciprocal overlap criterion for sensitivity assessment (which required SV sizes inferred on different experimental platforms to be in close agreement), our sensitivity estimates decreased by 12% and 18%, respectively, in trio and low-coverage sequence (Supplementary Table 8). We examined further an alternative SV discovery approach that involved the pairwise integration of deletion discovery methods, and tested its ability to discover SVs without relying on the inclusion of lower specificity calls following experimental validation (this approach resulted in the generation of the 'algorithm-centric set'; Fig. 1b). Whereas this alternative approach resulted in an increased number (by ~13%) of high-specificity (FDR < 10%) calls compared to the release set (Supplementary Text), overall it resulted in fewer SV calls owing to its decreased sensitivity at the lower (<200 bp) SV size range. In the following analyses we thus focused on the release set.

Extent and impact of our SV discovery set

We next assessed the extent and impact of our SV discovery (release) set. The median SV size was 729 bp (mean = 8 kilobases), approximately four times smaller than in a recent tiling CGH-based study[1], reflecting the high resolution of DNA-sequence-based SV discovery. We also compared our set to a recent survey of SVs in an individual genome[36] based on capillary sequencing and array-based analyses[24], and observed a similar size distribution for deletions, but differences in the size distributions of other SV classes, reflecting underlying differences in SV ascertainment (Supplementary Fig. 6). By comparing our SVs to databases of structural variation and to additional personal genome data sets, we classified 15,556 SVs in our set as novel, with an enrichment of low frequency SVs and small SVs amongst the novel variants (Methods and Supplementary Text).

A major advantage of sequence-based SV discovery is the nucleotide resolution mapping of SVs. We initially mapped the breakpoints of 7,066 deletions and 3,299 MEIs using SR and AS features. Using the TIGRA-targeted assembly approach (Chen, L. *et al.*, unpublished results) we further identified the breakpoints of an additional 4,188 deletions and 160 tandem duplications, initially discovered by RD, RP and PD methods (Methods, Supplementary Tables 3, 4). Altogether, we mapped ~15,000 SVs at nucleotide resolution, 48% of which were novel. Few deletion loci (4.4%) displayed different SV breakpoints in different samples, which is explainable by rare TIGRA misassemblies, or alternatively, by recurrently formed, multi-allelic SVs (Supplementary Text). TIGRA further enabled us to validate an additional 7,359 SVs by identifying the SVs' breakpoints (Methods), and to evaluate the mapping precision of SV discovery methods (Fig. 2c, Supplementary Fig. 2).

We assessed the putative functional impact of SVs in our set further by relating them to genomic annotation. Many SVs (1,775) affected coding sequences, resulting in full gene overlaps or exon disruptions (Table 2), many of which led to out-of-frame exons (Supplementary

Table 1 | Summary of discovered structural variation

	Deletions	Tandem duplications	Mobile element insertions	Novel sequence insertions	Total
Individual call-sets < 10% FDR	11,215	501	5,371	–	17,087
Validated experimentally*	10,810	–	–	128	10,938
Release set	22,025	501	5,371	128	28,025

* Only tabulates validated calls which were not already present in individual call-sets with less than 10% FDR.

Table 2 | Functional impact of our fine resolution SV set

SV class	Gene overlap				Total gene overlap	Total intergenic
	Full gene overlap	Coding exon affected, partial	UTR overlap	Intron overlap		
Deletions	654 (631)	1,093 (1,031)	315 (290)	7,319 (6,481)	9,381 (8,433)	12,644 (10,386)
Tandem duplications	2 (2)	7 (6)	9 (5)	197 (62)	215 (75)	286 (76)
Mobile element insertions	–	3 (2)	36 (26)	1,304 (97)	1,348 (112)	4,023 (758)
Novel sequence insertions	–	–	2 (2)	49 (49)	51 (51)	77 (77)
Sum	656 (633)	1,119 (1,040)	351 (309)	8,869 (6,689)	10,995 (8,671)	17,030 (11,280)

Figures in parentheses indicate numbers of validated SVs per category. We inferred gene overlap with Gencode gene annotation[44].

Table 9). We related gene disruptions to gene functions, and observed significant enrichments for several functional categories, including cell defence and sensory perception (Supplementary Table 10). High levels of structural variation, including copy number variation, were described previously for both processes[15,22,37]. These SVs might be maintained in the population by selection for the purpose of functional redundancy. Whereas most SVs intersecting with genes were deletions, several validated tandem duplications and MEIs also intersected with coding sequences (Table 2).

Population genetic properties of deletions

We next sought to generate genotypes for deletions discovered in the 1000GP data, both to facilitate population genetics analyses and to make our SV set amenable to association studies in the form of a reference genotype set. In this regard, the Genome STRiP genotyping method was developed (Handsaker, R. E., Korn, J. M., Nemesh, J. and McCarroll, S. A., unpublished results), a method combining information from RD, RP, SR and haplotype features of population-scale sequence data for genotyping (Methods, Supplementary Text). Using this approach we generated genotypes for 13,826 autosomal deletions in 156 individuals. The genotypes displayed 99.1% concordance with CGH array-based[1] genotypes (available for 1,970 of the deletions), indicating high genotyping accuracy.

Figure 3 presents allele frequency analyses based on these genotypes. As expected, common polymorphisms (minor allele frequency (MAF) > 5%) were typically shared across populations, whereas rare alleles were frequently observed in only one population (Fig. 3a–c). We observed several candidates for monomorphic deletions (that is, genomic segments putatively deleted in all individuals), explainable by rare insertions present in the reference genome or by remaining genotyping inaccuracies (Supplementary Text).

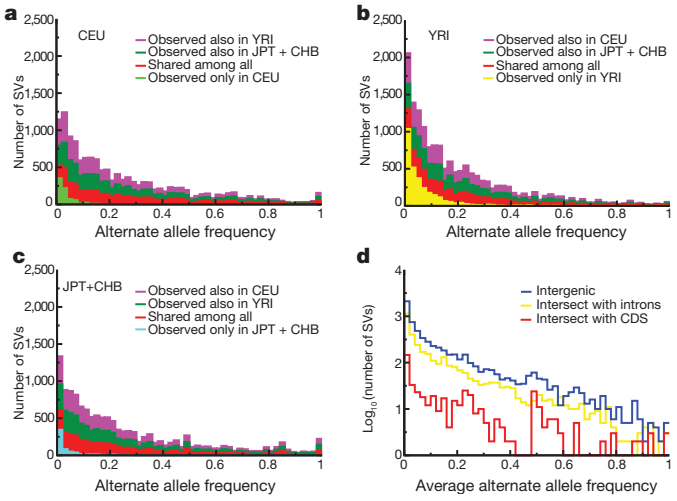

Figure 3 | Analysis of deletion presence and absence in three populations. **a**–**c**, Deletion allele frequencies and observed sharing of alleles across populations, displayed for deletions discovered in the CEU (**a**), YRI (**b**) and JPT+CHB (**c**) population samples in terms of stacked bars. **d**, Allele frequency spectra for deletions intersecting with intergenic (blue), intronic (yellow) and protein-coding sequences (red).

Next we assessed the allele frequencies of gene deletions. Similar to a recent array-based study[1], we observed a depletion of high-frequency alleles among deletions intersecting with protein-coding sequence compared to other deletions ($P = 2.2 \times 10^{-16}$; KS test), consistent with purifying selection keeping most gene deletions at low frequency (Fig. 3d). Nonetheless, several coding sequence deletions were observed with high allele frequency (>80%). Most of these occurred in regions annotated as segmental duplications, consistent with lessened evolutionary constraint in functionally redundant gene categories[22]. Intriguingly, common gene deletions also affected many unique genes with no obvious paralogues. We further analysed the abundance of gene deletions in different populations and observed highly differentiated loci, albeit with no statistically significant relationship between differentiation and particular categories of gene overlap, that is, intronic versus exonic (Supplementary Text).

By comparing deletion genotypes with genotypes of nearby SNPs, we found, consistent with earlier studies[1,13,38], that deletions in genomic regions accessible to short read sequencing display extensive linkage disequilibrium (LD) with SNPs. Most common deletions (81%) had one or more SNPs with which they are strongly correlated ($r^2 > 0.8$; Supplementary Fig. 7). This indicates that many deletions mapped in our study will be identifiable through tagging SNPs in future studies (Supplementary Text). On the other hand, a fifth of the genotyped deletions were not tagged by HapMap SNPs (a figure similar to the fraction of SNPs that are not tagged by HapMap SNPs[39]), implying that these SVs should be genotyped directly in association studies. Furthermore, the LD properties of complex SVs (for example, multiallelic SVs) have not yet been fully ascertained as methods for genotyping such SVs with similar accuracy are still being developed.

SV formation mechanism analysis

Nucleotide resolution breakpoint information enables inference of SV formation mechanisms[15,22]. Recent studies broadly distinguished between several germline rearrangement classes, some of which may comprise more than one SV formation mechanism[15,22,40,41]: non-allelic homologous recombination (NAHR), associated with long sequence similarity stretches around the breakpoints; rearrangements in the absence of extended sequence similarity (abbreviated as 'non-homologous' or NH), associated with DNA repair by non-homologous end-joining (NHEJ) or with microhomology-mediated break-induced replication (MMBIR); the shrinking or expansion of variable number of tandem repeats (VNTRs), frequently involving simple sequences, by slippage; and MEIs. We distinguished among the classes NAHR, NH, VNTR and MEI by examining the breakpoint junction sequences of SVs that had initially been discovered as deletions or tandem duplications relative to a human reference.

We first compared these SVs to orthologous primate genomic regions to distinguish deletions from insertions/duplications with respect to reconstructed ancestral loci using the BreakSeq classification approach[41]. This analysis showed that of the 11,254 nucleotide-resolution SVs discovered as deletions relative to a human reference, 21% actually represented insertions and 2% represented tandem duplications relative to the putative ancestral genome. Of the remaining SVs, 60% were classified as deletions relative to ancestral sequence, whereas the ancestral state of 17% was undetermined. By comparison,

out of 160 nucleotide-resolution SVs identified as tandem duplications relative to the reference genome, 91.6% were classified as duplications relative to the ancestral genome, whereas the ancestral state of 8.4% remained undetermined (Supplementary Text). Our breakpoint analysis revealed that 70.8% of the deletions and 89.6% of the insertions exhibited breakpoint microhomology/homology ranging from 2–376 bp in size, with distribution modes of 2 bp (attributable to NH) and 15 bp (attributable to MEI), respectively (Fig. 4a, Supplementary Text). As expected[40], a small portion of the deletions (16.1%) displayed non-template inserted sequences at their breakpoint junctions. By comparison, the tandem duplications showed extensive stretches displaying ≥95% sequence identity at the breakpoints linearly correlating in length with SV size (Fig. 4a). In addition, most tandem duplications displayed 2–17 bp of microhomology at the breakpoint junctions (Supplementary Text).

We subsequently applied BreakSeq[41] to infer formation mechanisms for all SVs classified with regard to ancestral state. Using BreakSeq, we inferred NH as the dominating deletion mechanism, and MEI as the dominating insertion mechanism (Fig. 4b, c and Supplementary Table 11). Furthermore, an abundance of microhomology at tandem duplication breakpoints suggested frequent formation of this SV class by a rearrangement process acting in the absence of homology (NH). When relating SV formation to the variant size spectrum, we observed marked insertion peaks for MEIs at 300 bp, corresponding to *Alu* elements, and at 6 kb, corresponding to the L1 class of long interspersed elements (LINEs) (Fig. 4c). By comparison, NH- and NAHR-based mechanisms occurred across a wide size-range, whereas VNTR expansion/shrinkage, consistent with earlier findings[1], led to relatively small SV sizes (Fig. 4c, d).

Furthermore, when displaying the genomic distribution of SVs (Fig. 5a), we observed a notable clustering of SVs into 'SV hotspots'. We analysed this clustering in detail by examining the distribution of non-overlapping, adjacent SVs, and observed a marked clustering of SVs formed by NAHR, VNTR and NH, respectively, a signal extending to hundreds of kilobases (Fig. 5b). The clustering was influenced by an abundance of VNTR near the centromeres[41] and NAHR near the telomeres (Fig. 5a). A significant enrichment of NAHR near recombination hotspots ($P = 1.3 \times 10^{-15}$) and segmental duplications ($P = 3.1 \times 10^{-17}$) further contributed to the clustering (Supplementary Table 13).

To further explore this clustering we devised a segmentation approach for predicting SV hotspots (Methods), which yielded a map of 51 putative SV hotspots (Supplementary Table 14). Most of the hotspots (80%) mainly comprised SVs originating from a single formation mechanism (Fig. 5c). Most of these corresponded to NAHR hotspots, although hotspots dominated by NH and VNTR were also evident. These observations indicate that SV formation is frequently associated with the locus-specific propensity for genomic rearrangement.

Conclusions and discussion

By generating an SV set of unprecedented size along with breakpoint assemblies and reference genotypes, we demonstrate the suitability of population-scale sequencing for SV analysis. Nucleotide resolution data allow the construction of reference data sets and make SVs readily assessable across different experimental platforms using genotyping approaches. Our fine-scale map enabled us to examine the functional impact of SVs, as exemplified by the set of gene disruption variants we reported, which will be of value for genome and exome sequencing studies.

Our map further enabled us to examine size spectra of SV formation mechanisms and led us to identify genomic SV hotspots that are commonly dominated by a single formation mechanism. Recurrent rearrangements, implicated in genomic disorders, are hypothesized to be associated with local genome architecture[42], for example, with segmental duplications that facilitate NAHR. Also, DNA rearrangement in the absence of homology, that is, MMBIR, has been implicated

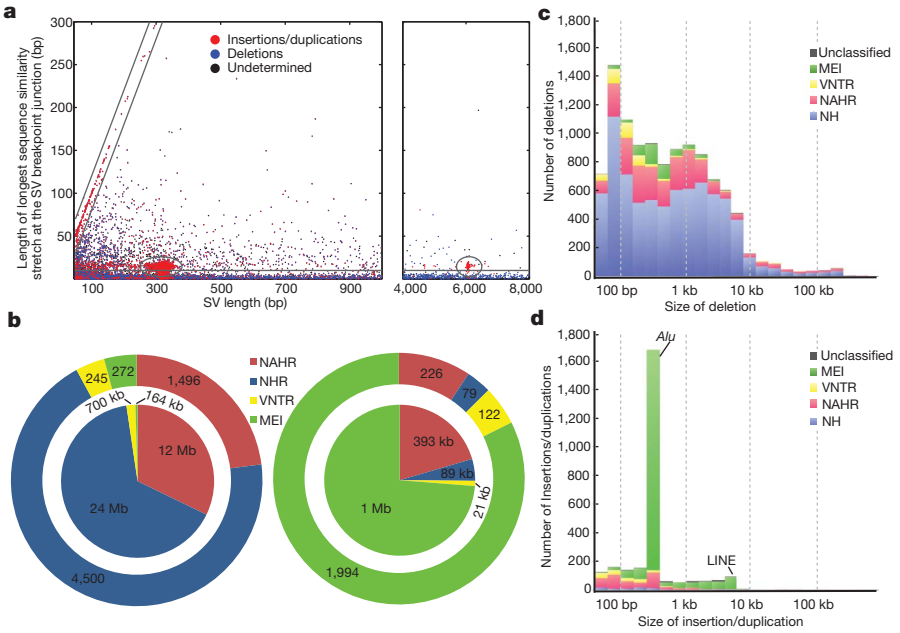

Figure 4 | Contribution of SV formation mechanisms to the SV size spectrum. a, Breakpoint junction homology/microhomology length plotted as a function of SV size for SVs originally identified as deletions compared to a human reference. Dots are coloured according to the SVs' classification as deletions, insertions/duplications, or 'undetermined' relative to inferred ancestral genomic loci. Gray lines mark groups of SVs likely formed by a common formation mechanism. The diagonal highlights tandem duplications (and few reciprocal deletion events), in which the length of the duplicated sequence correlates linearly with the length of the longest breakpoint junction sequence identity stretch. The ellipses indicate MEIs, that is, *Alu* (~300 bp) and L1 (~6 kb) insertions, associated with target site duplications of up to 28 bp in size at the breakpoints. The horizontal group corresponds mostly to NH-associated deletions with <10 bp microhomology at the breakpoints. The remaining (ungrouped) SVs comprise truncated MEIs, VNTR expansion and shrinkage events, as well as NAHR-associated deletions and duplications. **b**, Relative contributions of SV formation mechanisms in the genome. Numbers of SVs are displayed on the outer pie chart and affected base pairs on the inner. Left panel, SVs classified as deletions relative to ancestral loci. Right panel, SVs classified as insertions/duplications. **c**, Size spectra of deletions classified relative to ancestral loci. **d**, Size spectra of insertions/duplications.

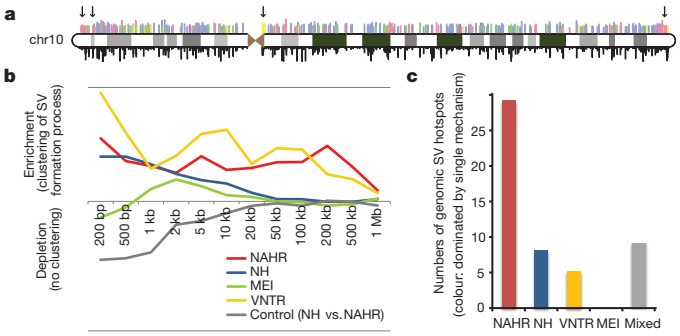

Figure 5 | Mapping hotspots of SV formation in the genome. **a**, Distribution of SVs on chromosome 10 ('chr10'). Above the ideogram, coloured bars indicate SV formation mechanisms (same colour scheme as in (**b**) and (**c**)); bar lengths relate to the logarithm of SV size. Below the ideogram, bar lengths are directly proportional to allele frequencies. Arrows indicate an SV hotspot near the centromere underlying mainly VNTR and several hotspots near the telomeres underlying mainly NAHR events. **b**, Enrichment of SVs inferred to be formed by the same formation mechanism for different genomic window sizes. Displayed is an enrichment of nearby, non-overlapping SVs formed by the same mechanism relative to an SV set where mechanism assignments are shuffled randomly. **c**, SV hotspots are mostly dominated by a single formation mechanism. Coloured bars depict numbers of SV hotspots in which at least 50% of the variants were inferred to be formed by a single formation mechanism. The average abundance of NAHR-classified SVs in NAHR hotspots was 70% (compared with 77% for VNTR-hotspots; 69% for NH). The grey bar ('mixed') corresponds to SV hotspots with no single mechanism dominating.

in recurrent SV formation[8,43]. In this regard, we noticed that out of the hotspots we report, six fall into critical regions of known genetic disorders associated with recurrent *de novo* deletions, including Miller–Dieker syndrome and Leri–Weill dyschondrosteosis (Supplementary Table 14). Irrespective of potential disease relevance, or inferred mechanism of formation, our analysis revealed a map of SV hotspots that may constitute local centres of *de novo* SV formation, consistent with the concept that local genome architecture contributes to genomic instability[42].

Our study focused on characterizing deletions, which are often associated with disease[9]. Facilitated by ancestral analyses of SV loci, we also characterized insertions and tandem duplications, albeit in less detail than deletions. Companion papers with more detailed analyses of MEIs and copy number variation within segmental duplications are published elsewhere (Stewart, C. *et al.*, unpublished results, and ref. 34). Of note, most SV discovery methods depend on mapping reads onto their genomic locus of origin, that is, the 'accessible' fraction of the genome, a fraction lessened in segmental duplications that are of high interest to SV analysis. Nonetheless, owing to the abilities of SV discovery methods in detecting SVs in these regions and in interpreting reads with multiple mapping positions, the 'accessible' fraction of the genome is higher for SVs than for SNPs[16]. In the future, sequencing technologies generating longer DNA reads will increase the accessible genome, and will enable the assessment of SVs embedded in long repeat structures, such as balanced inversions.

Our SV resource will enable the discovery, genotyping and imputation of SVs in larger cohorts. Numerous genomes will be sequenced in the coming months to facilitate disease association studies. Systematic characterization of SVs in these genomes will benefit from the concepts and data sets presented here.

METHODS SUMMARY

Samples. Whole genome sequencing data for 179 unrelated individuals and six individuals from parent-offspring trios were obtained as part of the 1000GP. These data were generated with Illumina/Solexa, Roche/454 and Life Technologies/SOLiD sequencing technology platforms.

SV discovery and breakpoint assembly. The SV discovery methods we applied comprised six RP, four RD, three SR, four AS, and two PD based methods. TIGRA (Chen, L. *et al.*, unpublished results) was used for targeted breakpoint assembly.

Experimental validation. We validated SV calls by PCR, array CGH and SNP microarrays, targeted assembly, and custom microarray-based sequence capture. PCR was performed in various different laboratories[33], CGH analysis was performed based on tiling array data provided by the Genome Structural Variation Consortium (ArrayExpress: E-MTAB-40), and SNP array analysis based on data obtained from the International HapMap Consortium (http://hapmap.ncbi.nlm.nih.gov).

Genotyping. Genome STRiP (Handsaker, R. E., Korn, J. M., Nemesh, J. and McCarroll, S. A., unpublished results) was used for deletion genotyping in low-coverage sequence data. Initial genotype likelihoods were derived with a Bayesian model and imputation into a SNP genotype reference panel from the HapMap[39] (Hapmap3r2) was achieved with Beagle (v3.1; http://faculty.washington.edu/browning/beagle/beagle.html).

SV formation mechanism analysis. SV breakpoints mapped at nucleotide resolution were analysed with BreakSeq[41] to classify SVs relative to putative ancestral loci and to infer SV formation mechanisms. SV hotspots were mapped with custom Perl and R scripts.

Received 19 August; accepted 26 November 2010.

1. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464,** 704–712 (2010).
2. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466,** 368–372 (2010).
3. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316,** 445–449 (2007).
4. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455,** 232–236 (2008).
5. McCarthy, S. E. *et al.* Microduplications of 16p11.2 are associated with schizophrenia. *Nature Genet.* **41,** 1223–1227 (2009).
6. Craddock, N. *et al.* Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464,** 713–720 (2010).
7. McCarroll, S. A. *et al.* Deletion polymorphism upstream of *IRGM* associated with altered *IRGM* expression and Crohn's disease. *Nature Genet.* **40,** 1107–1112 (2008).
8. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nature Rev. Genet.* **10,** 551–564 (2009).
9. Stankiewicz, P. & Lupski, J. R. Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* **61,** 437–455 (2010).
10. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305,** 525–528 (2004).
11. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nature Genet.* **36,** 949–951 (2004).
12. Sharp, A. J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77,** 78–88 (2005).
13. McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genet.* **40,** 1166–1174 (2008).
14. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nature Genet.* **37,** 727–732 (2005).
15. Korbel, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318,** 420–426 (2007).
16. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genet.* **41,** 1061–1067 (2009).
17. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods* **6,** 677–681 (2009).
18. 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 (2009).
19. Medvedev, P., Stanciu, M. & Brudno, M. Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods* **6,** S13–S20 (2009).
20. 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 (2009).
21. Chiang, D. Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature Methods* **6,** 99–103 (2009).
22. Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453,** 56–64 (2008).
23. Lee, S., Cheran, E. & Brudno, M. A robust framework for detecting structural variations in a genome. *Bioinformatics* **24,** i59–i67 (2008).
24. Pang, A. W. *et al.* Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.* **11,** R52 (2010).
25. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297,** 1003–1007 (2002).
26. Campbell, P. J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genet.* **40,** 722–729 (2008).
27. 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 (2009).
28. Mills, R. E. *et al.* An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* **16,** 1182–1190 (2006).

29. 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 (2009).
30. Simpson, J. T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19,** 1117–1123 (2009).
31. Hajirasouliha, I. *et al.* Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics* **26,** 1277–1283 (2010).
32. Li, R. *et al.* The sequence and *de novo* assembly of the giant panda genome. *Nature* **463,** 311–317 (2010).
33. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
34. Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330,** 641–646 (2010).
35. Willer, C. J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature Genet.* **41,** 25–34 (2008).
36. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5,** e254 (2007).
37. Hasin-Brumshtein, Y., Lancet, D. & Olender, T. Human olfaction: from genomic variation to phenotypic diversity. *Trends Genet.* **25,** 178–184 (2009).
38. Hinds, D. A., Kloek, A. P., Jen, M., Chen, X. & Frazer, K. A. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nature Genet.* **38,** 82–85 (2006).
39. Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).
40. Conrad, D. F. *et al.* Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nature Genet.* **42,** 385–391 (2010).
41. Lam, H. Y. *et al.* Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nature Biotechnol.* **28,** 47–55 (2010).
42. Lupski, J. R. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* **14,** 417–422 (1998).
43. Lee, J. A., Carvalho, C. M. & Lupski, J. R. A. DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131,** 1235–1247 (2007).
44. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* **7** (suppl. 1), S4 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We would like to acknowledge C. Hardy, R. Smith, A. De Witte and S. Giles for their assistance with validation. M.A.B.'s group was supported by a grant from the National Institutes of Health (RO1 GM59290) and G.T.M.'s group by grants R01 HG004719 and RC2 HG005552, also from the NIH. J.O.K.'s group was supported by an Emmy Noether Fellowship of the German Research Foundation (Deutsche Forschungsgemeinschaft). J.W.'s group was supported by the National Basic Research Program of China (973 program no. 2011CB809200), the National Natural Science Foundation of China (30725008; 30890032; 30811130531; 30221004), the Chinese 863 program (2006AA02Z177; 2006AA02Z334; 2006AA02A302; 2009AA022707), the Shenzhen Municipal Government of China (grants JC200903190767A; JC200903190772A; ZYC200903240076A; CXB200903110066A; ZYC200903240077A; ZYC200903240076A and ZYC200903240080A) and the Ole Rømer grant from the Danish Natural Science Research Council. E.E.E.'s group was supported by grants P01 HG004120 and U01 HG005209 from the National Institutes of Health. C.L.'s group was supported by grants from the National Institutes of Health: P41 HG004221, RO1 GM081533 and UO1 HG005209 and X.S. was supported by a T32 fellowship award from the NIH. We thank the Genome Structural Variation Consortium (http://www.sanger.ac.uk/humgen/cnv/42mio/) and the International HapMap Consortium for making available microarray data. The authors acknowledge the individuals participating in the 1000 Genomes Project by providing samples, including the Yoruba people of Ibadan, Nigeria, the community at Beijing Normal University, the people of Tokyo, Japan, and the people of the Utah CEPH community. Furthermore, we thank R. Durbin and L. Steinmetz for comments on the manuscript.

Author Contributions The authors contributed this study at different levels, as described in the following. SV discovery: K.W., C.S., R.E.H., K.C., C.A., A.A., S.C.Y., R.K.C., A.C., Y.F., I.H., F.H., Z.I., D.K., R.Li., Y.L., C.L., R.Lu., X.J.M., H.E.P., L.D., G.T.M., J.S., Ju.W., Ka.Y., Ke.Y., E.E.E., M.B.G., M.E.H., S.A.M. and J.O.K. SV validation: R.E.M., K.W., K.C., A.A., S.C.Y., F.G., M.K.K., J.K., J.N., A.E.U., X.S., A.M.S., J.A.W., Y.Z., Z.D.Z., M.A.B., J.S., M.S., M.E.H., C.L. and J.O.K. SV genotyping: K.W., R.E.H., J.K., J.N., M.E.H. and S.A.M. Data analysis: R.E.M., C.S., C.A., A.A., R.E.H., K.C., S.C.Y., R.K.C., A.C., D.F.C., Y.F., F.H., L.M.I., Z.I., J.M.K., M.K.K., S.K., J.K., E.K., D.K., H.Y.K.L., J.L., R.Li, Y.L., C.L., R.Luo, X.J.M., J.N., H.E.P., T.R., A.S., X.S., M.P.S., J.A.W., Ji.W., Y.Z., Z.D.Z., M.A.B., L.D., G.T.M., G.M., J.S., M.S., Ju.W., Ka.Y., Ke.Y., E.E.E., M.B.G., M.E.H., C.L, S.A.M. and J.O.K. Preparation of manuscript display items: R.E.M., K.W., C.S., C.A., A.A., R.E.H., S.C.Y., L.M.I., S.K., E.K., M.K.K., X.J.M., X.S., J.A.W., M.B.G., S.A.M. and J.O.K. Co-chairs of the Structural Variation Analysis group: E.E.E., M.E.H. and C.L. The following equally contributed to directing the described analyses and participating in the design of the study and should be considered joint senior authors: E.E.E., M.B.G., M.E.H., C.L., S.A.M. and J.O.K. The manuscript was written by the following authors: R.E.M. and J.O.K.

Author Information Data sets described here can be obtained from the 1000 Genomes Project website at www.1000genomes.org (July 2010 Data Release). Individual SV discovery methods can be obtained from sources mentioned in Supplementary Table 2, or upon request from the authors. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.O.K. (jan.korbel@embl.de).