Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome *de novo* assembly

Yingrui Li[1,10], Hancheng Zheng[1,10], Ruibang Luo[1–3,10], Honglong Wu[1,4,10], Hongmei Zhu[1], Ruiqiang Li[1], Hongzhi Cao[1,4], Boxin Wu[1], Shujia Huang[1,2], Haojing Shao[1,2], Hanzhou Ma[1,2], Fan Zhang[1,2], Shuijian Feng[1], Wei Zhang[1], Hongli Du[2], Geng Tian[1], Jingxiang Li[1], Xiuqing Zhang[1], Songgang Li[1], Lars Bolund[1,5], Karsten Kristiansen[1,6], Adam J de Smith[7], Alexandra I F Blakemore[7], Lachlan J M Coin[8], Huanming Yang[1], Jian Wang[1] & Jun Wang[1,6,9]

Here we use whole-genome *de novo* assembly of second-generation sequencing reads to map structural variation (SV) in an Asian genome and an African genome. Our approach identifies small- and intermediate-size homozygous variants (1–50 kb) including insertions, deletions, inversions and their precise breakpoints, and in contrast to other methods, can resolve complex rearrangements. In total, we identified 277,243 SVs ranging in length from 1–23 kb. Validation using computational and experimental methods suggests that we achieve overall <6% false-positive rate and <10% false-negative rate in genomic regions that can be assembled, which outperforms other methods. Analysis of the SVs in the genomes of 106 individuals sequenced as part of the 1000 Genomes Project suggests that SVs account for a greater fraction of the diversity between individuals than do single-nucleotide polymorphisms (SNPs). These findings demonstrate that whole-genome *de novo* assembly is a feasible approach to deriving more comprehensive maps of genetic variation.

The completion of the International Human Genome Project[1–3] expedited, and in some cases made possible, the identification of genetic variations for tracing evolution, determining population patterns[4,5] and assessing disease susceptibility[6–9] as well as other phenotypic traits[10]. The International HapMap project[11] and genome-wide association studies (GWAS)[12,13] have allowed extensive research based on characterization of single-nucleotide polymorphisms (SNPs). This was followed by work to identify and characterize structural variations, including insertions, deletions, inversions and other DNA sequence rearrangements. A large number of such variations have been discovered in the human genome putatively having equal or greater functional impact than SNPs[14–23]. These variations have been identified using either (i) optical signals from array-based technologies, (ii) aberrant read depth, (iii) gapped alignment of reads or so-called split-read methods, (iv) deviation from mean library insert-size in paired-end mapping and (v) gapped alignment of contigs assembled from 'semi-aligned' paired-end reads (that is, one read is too different from the reference to be confidently aligned as the other side could be well aligned, which enables the read to be uniquely anchored).

Existing methods for calling structural variations from short sequencing reads are hampered by one or more of the following limitations: (i) the methods may favor a particular length range of structural variations; (ii) they may favor discovery of particular types of structural variations; (iii) they may be unable to resolve the exact structural variation genotypes and/or breakpoints at single nucleotide resolution; and (iv) because of difficulties mapping reads to the genome, they may not be able to accurately identify complex rearrangements. Paired-end mapping, for example, can only predict insertion breakpoints within a few base pairs of the exact breakpoint position[24], and it can only detect insertions when the entire sequence is contained within the DNA fragment whose ends are being sequenced; thus, the maximum size of an insertion that can be detected by paired-end mapping is limited by the largest insert size present in a library. Split-read methods, on the other hand, can precisely define a breakpoint and genotype of an insertion, but only when it is shorter than the read length. Thus, studies carried out so far have been of limited completeness, accuracy and/or resolution. A recent pipeline has been developed to integrate methods for structural variation screening and applied local assembly to validate and recover the breakpoints[25], which introduces the intriguing possibility of directly ascertaining structural variations with breakpoints from *de novo* assembly of the whole genome.

In theory, accurate and complete *de novo* assembly of human genomes should allow relatively more comprehensive mapping of structural variations[26]. Until now, the cost of conventional Sanger sequencing and the difficulty of assembling data from massively parallel

[1]BGI-Shenzhen, Shenzhen, China. [2]School of Bioscience and Biotechnology, South China University of Technology, Guangzhou, China. [3]Department of Computer Science, The University of Hong Kong, Hong Kong, China. [4]Genome Research Institute, Shenzhen University Medical School, Shenzhen, China. [5]Institute of Human Genetics, University of Aarhus, Aarhus, Denmark. [6]Department of Biology, University of Copenhagen, Copenhagen, Denmark. [7]Department of Genomics of Common Disease, School of Public Health, Imperial College London, London, UK. [8]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College, London, UK. [9]The Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark. [10]These authors contributed equally to this work. Correspondence should be addressed to J.W. (wangj@genomics.org.cn).

Published online 24 July 2011; doi:10.1038/nbt.1904

fragment sequencing data have restricted the practical use of this approach. However, the recent availability of large-scale genome assembly data from next-generation sequencing technologies, including new assembly algorithms[27–30], now allow researchers to develop more detailed structural variation maps for multiple *de novo* assemblies of human genomes at a lower cost.

Here we describe an approach that complements previous methods for reliable homozygous structural variation identification. Our approach accurately determines genotype and breakpoints relative to a reference genome based on *de novo* assembly of Illumina Genome Analyzer sequencing data. In this paper, we examined only homozygous structural variations because detecting heterozygous structural variations requires assembly of haplotype sequences, which is not yet possible using existing assemblers. Simulation and experimental validation demonstrated that this method produced detailed structural variation maps and allowed comprehensive characterization of structural variations in two human genomes. The two human structural variation maps were then used to study the genome-wide distribution and patterns of structural variation events in different categories. The results supported previous observations on characteristics and potential biological impacts of structural variation events. Finally, we also profiled the occurrences of structural variations identified in the two genomes in a population of 106 individuals from the 1000 Genomes Project[31] to infer structural variation frequency distributions in the human population. These analyses indicate that structural variations may in general be undergoing stronger negative selection compared with SNPs, and thus be more likely to be identified as unique to an individual than SNPs.

RESULTS

Structural variation detection in short-read whole-genome assemblies

To detect structural variations, we used LASTZ[32], a replacement for BLASTZ[33] that is optimized for aligning whole genomes, to align the *de novo* assemblies[34] of an Asian (YH; scaffold N50, 446.3 kb; contig N50, 7.4 kb) and an African genome (NA18507, NCBI Short Read Archive accession number SRA000271; scaffold N50, 61.9 kb; contig N50, 5.9 kb) onto the NCBI human reference genome build 36 (NCBI36, http://www.ncbi.nlm.nih.gov). We limited our analysis to structural variations <50 kb, as larger variations can be detected by traditional approaches including read depth and array-based technologies, and they are computationally very costly to find using our method because of limitations of the LASTZ alignment algorithm. To identify and remove errors in pair-wise alignments, we developed a dynamic programming algorithm that prefers syntenic alignment and ensures only one best alignment result for each locus in the reference genome. This algorithm is designed to eliminate the numerous errors that occur when doing genome-to-genome pair-wise alignment and to guarantee the co-linearization of the alignment (Online Methods). Then, we extracted gaps and segmental rearrangements in pair-wise alignments as candidate structural variation loci; these variations include genotype and breakpoint information (**Fig. 1a** and **Supplementary Fig. 1**).

We next developed filters to eliminate spurious structural variation calls (false positives). First, we computed read coverage at each structural variation locus by aligning the Illumina Genome Analyzer sequencing reads (GA reads) onto the NCBI36 and the two assemblies using the BWA tool[35] with the gapped alignment option enabled. Because of limitations of BWA[35], we used different filtering approaches for candidate structural variation ≤50 bp (set 1) or >50 bp (set 2). For structural variations ≤50 bp, we identified false positives as those supported by fewer than four gapped aligned reads or if any reads were aligned inconsistently with the breakpoints and/or genotype predicted from mapping the assembly to the reference. For structural variations >50 bp, we reasoned that false positives could be identified by inconsistencies in paired-end and read-depth data near to the putative breakpoints. For reads aligned to the *de novo* assembly, the expectation is that an authentic structural variation would be covered by sufficient paired-end reads with proper span size and strand orientation, whereas spurious structural variations would be covered by abnormally mapped read pairs that could only be aligned as two independent single-end reads around the putative breakpoints. For reads aligned to the reference genome, the opposite would be expected. In addition, the overall depth of paired-end mapping across a true insertion region on the assembly should be more consistent than for a false one, and a true deletion on the reference should have an overall lower depth than the average depth of the whole reference genome (**Supplementary Fig. 2**). Based on this logic, we devised a metric, the S/P ratio, to quantify the reliability of structural variation candidates >50 bp. The S/P ratio is computed by dividing the number of successfully aligned single-end reads by the number of aligned paired-end reads at a structural variation locus. We compute S/P ratios for each structural variation locus in both the *de novo* assembly and reference, and use Fisher's exact test to determine whether there is a significant difference between the two. Spurious structural variations will have similar S/P ratios in the *de novo* assembly and the reference.

After filtering out false-positive structural variations called for the Asian and African genome assemblies, we identified 80,719 and 87,457 insertions (ranging from 1–23,203 bp in length), 51,711 and 56,074 deletions (1–7,916 bp in length), 26 and 23 inversions (10–21,052 bp in length) and 717 and 516 complex rearrangements (defined as intra- or interchromosome duplications and translocations, 100–5,683 bp in length) in the two genomes, respectively (**Table 1** and **Fig. 1b**,**c**). Notably, our method identifies the precise breakpoints and genotypes of all these structural variations.

The distribution of the lengths of the structural variations are consistent with previous findings[19,20,36] in that longer variations were less abundant. The only exception was the small increase in the number of structural variations in the size range of 200–400 bp; this was due to the enrichment of *Alu* element insertions and deletions (**Fig. 1d**,**e**), as demonstrated in previous studies[19]. Our analysis showed that insertions and deletions have length distribution peaks in coding sequences at positions that are multiples of three owing to negative selection of frame-shift indels[37] (**Fig. 1f**). This finding also indicated that our structural variation calls of small indels were sufficiently accurate to eliminate background noise caused by spurious indels when defining structural variation patterns within different genomic features, such as coding regions.

We also examined complex rearrangement cases that are generally the result of an intricate combination of several insertions, deletions and/or inversions. As our structural variation calls were directly ascertained from assembled sequences, we were able to resolve complex rearrangements. **Supplementary Figure 3** shows an example of such a region with complex rearrangements that usually cannot be resolved by methods based on paired-end mapping, read depth or split reads.

Precision and sensitivity of structural variant calls

To assess the accuracy of our detection method, we simulated homozygous indels at random sites throughout chromosome 17 of the HuRef assembly (http://huref.jcvi.org/). We first introduced 16,490 indels ranging in length from 1 to 50 kbp, with adjustments made based on the distribution of sizes found previously in the two

genomes[20], into chromosome 17 of the HuRef assembly. Then, we simulated Illumina reads using the modified sequence (Online Methods), and identified structural variations using our method. The assembled scaffolds covered the flanking regions of 12,516 (75.9%) simulated indels (detectable indels). The remaining simulated indels (24.1%) were in regions of the genome that could not be assembled. In total, we found 11,311 (90.4%) of the detectable indels using our method. Of these, our method accurately identified the exact breakpoints and genotype for 11,194 (98.8%), while calling only 137 false positives.

Based on these results, we estimate that the false-positive rate of our method is 1.2%, and the false-negative rate is 9.6% in assembled regions. False positives were largely attributable to assembly errors or the presence of large repetitive elements. For example, analysis of the flanking sequences of the false-positive indels using RepeatMasker[38] revealed that 61.9% of these contained highly repetitive elements. Compared with an approach based on paired-end mapping, such as BreakDancer, which ascertained 74% of simulated indels >20 bp and 68% of simulated indels 10–20 bp long with a false-positive rate of ~10%, our method demonstrates similar sensitivity but improved precision (1 – false-negative rate) (**Table 2**).

We next used the simulated indels to assess the bias of our method towards detecting structural variations of different lengths. For small (1–10 bp), medium (10–50 bp) and large (>50 bp) indels, 20.4%, 29.8% and 29.2%, respectively, were not located in assembled sequences, largely because they were located in repetitive regions (**Fig. 2**). Thus, these indels were missed because of the limitations of current *de novo*

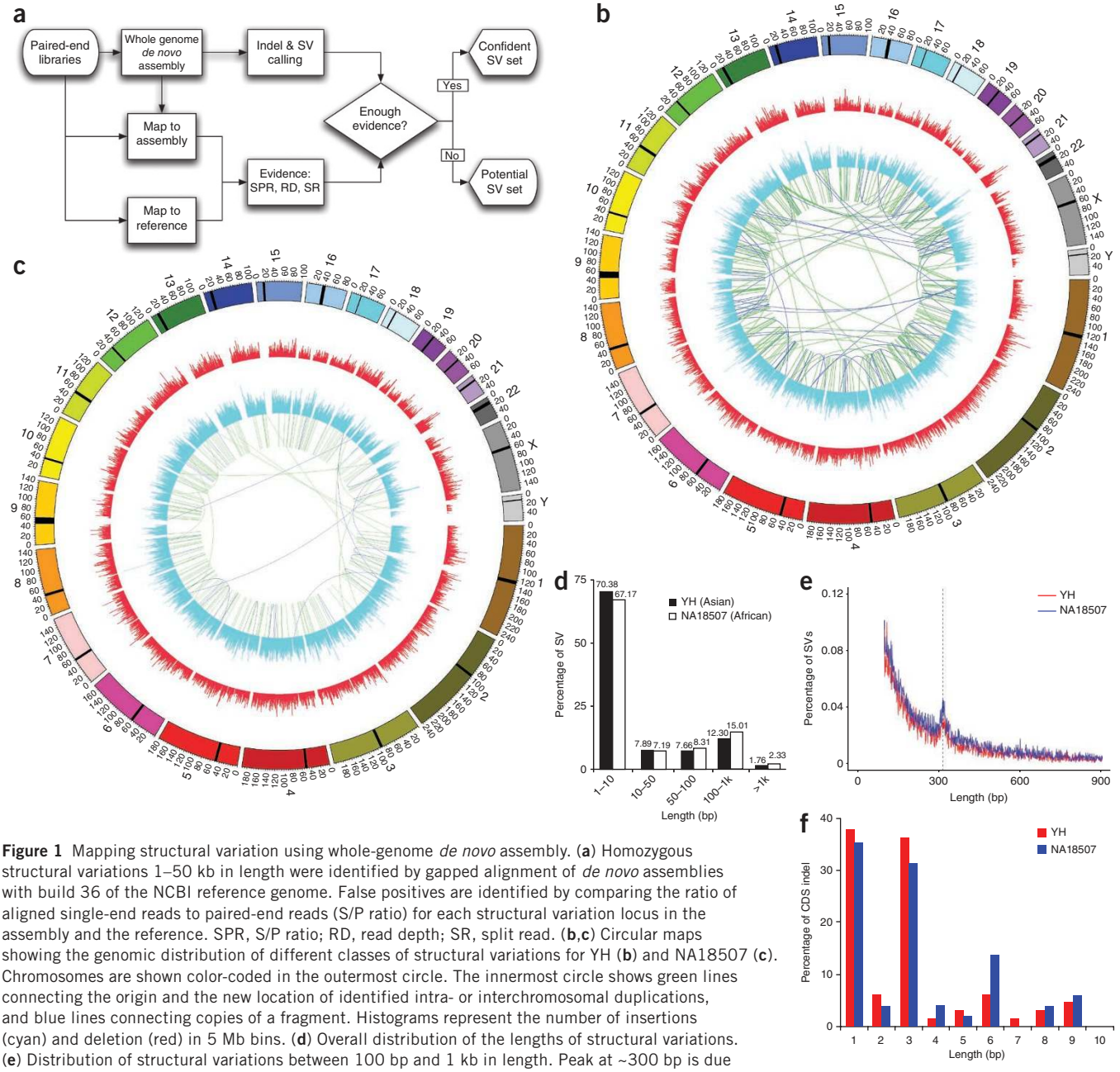

Figure 1 Mapping structural variation using whole-genome *de novo* assembly. (**a**) Homozygous structural variations 1–50 kb in length were identified by gapped alignment of *de novo* assemblies with build 36 of the NCBI reference genome. False positives are identified by comparing the ratio of aligned single-end reads to paired-end reads (S/P ratio) for each structural variation locus in the assembly and the reference. SPR, S/P ratio; RD, read depth; SR, split read. (**b**,**c**) Circular maps showing the genomic distribution of different classes of structural variations for YH (**b**) and NA18507 (**c**). Chromosomes are shown color-coded in the outermost circle. The innermost circle shows green lines connecting the origin and the new location of identified intra- or interchromosomal duplications, and blue lines connecting copies of a fragment. Histograms represent the number of insertions (cyan) and deletion (red) in 5 Mb bins. (**d**) Overall distribution of the lengths of structural variations. (**e**) Distribution of structural variations between 100 bp and 1 kb in length. Peak at ~300 bp is due to the enrichment of *Alu* element insertions and deletions. (**f**) Distribution of structural variations between 1 bp and 15 bp in length in coding sequences. Peaks at multiples of 3 bp can be explained by the fact that they are under weaker negative selection than are frame-shift indels with length not evenly divided by three[47].

Table 1 Summary of structural variations in YH and NA18507

		No. of confident SVs	Min. length (bp)	Max. length (bp)
NA18507	Insertion	87,457	1	22,617
	Deletion	56,074	1	7,916
	Inversion	23	53	13,149
	Complex	516	101	2,260
YH	Insertion	80,719	1	23,203
	Deletion	51,711	1	6,160
	Inversion	26	57	2,785
	Complex	717	100	5,683

We aligned the assembly of YH (Asian) and NA18507 (African) genome sequences against the NCBI human reference genome build 36 and refined the alignments to guarantee the accuracy and co-linearization of the results. Structural variations (SVs) were extracted from refined alignments. Confident SVs were those SVs that passed our filtering threshold as described in Online Methods.

assembly methods in resolving highly repetitive elements[39], rather than because of errors in the alignment, structural variation calling or filtering steps. In the assembled regions, medium-sized indels had a higher false-negative rate (19%) than small ones (6%), yet large indels were nearly all identified (false-negative rate <0.2%). The large indels had a false-positive rate of only 3%, which is approximately tenfold higher than the false-positive rates of other length ranges. In summary, our method enabled us to detect a more integrated spectrum of structural variation compared with previous approaches[40], although some biases in the discovery of structural variations with different lengths were observed.

Next, we assessed the accuracy of our method when applied to experimental data. We applied a number of different experimental approaches to generate high-accuracy sequences for validation purpose.

First, Sanger capillary sequencing was performed on 95 randomly selected structural variations from the YH genome that ranged from 1–50 bp (**Supplementary Table 1**). We consider validated structural variations to be those that had a sequence that exactly matched the detected structural variation sequence from the assembly. We were able to sequence 91 of the 95 structural variations and validate 88 (96.7%) of them.

Second, PCR sequencing was performed across the breakpoints of an additional 57 structural variations (40 insertions and 17 deletions) that were >2 kb. We successfully sequenced 29 of the insertions and all of the deletions. Of these, we validated 28 of the 29 insertions either at both breakpoints (14 of the insertions) or at one breakpoint (seven left breakpoints and seven right breakpoints), and 16 of 17 deletions were also confirmed to be correct.

Lastly, to check whether the 15 structural variations that we could not amplify by PCR (four indels that were <50 bp and 11 insertions >2 kb) were authentic or spurious, we sequenced a pooled fosmid library generated from YH genomic DNA using an Illumina GAIIx instrument. Each pool consisted of only 30 fosmids to avoid ambiguities owing to segmented duplications in the assembly process. And each pool was sequenced independently to reduce the complexity of assembling the reads and problems caused by repetitive sequences[41]. Using the contigs assembled from fosmids, we validated 11 (73%) of the remaining 15 structural variations, which indicates that failure to amplify a variant by PCR is not necessarily indicative of a spurious structural variation call (**Supplementary Table 2**).

In summary, 143 structural variations out of 152 (94.1%) structural variations were validated by experimental methods, which supports the accuracy of our methods on experimental data.

Performance comparison with other methods

We next benchmarked our method by comparing our predictions of structural variants in the NA18507 genome to predictions made using the bioinformatics tools BreakDancer[20], which uses paired-end read mapping, and pIndel[42], which is tailored to identify small insertions and deletions. Of the structural variations that we identified, 60.2% and

Table 2 Comparison between features of structural variation detection tools and benchmarking evaluations on structural variation calls between *de novo* assembly-based method, BreakDancer and pIndel

		SV detection tools		
		De novo assembly	BreakDancer	pIndel
Main detectable length range[a]		1 bp–50 kbp	>10 bp	1–10 kbp (deletions), 1–16 bp (insertions)
Detectable SV types				
Insertions		Yes	Yes	Yes
Deletions		Yes	Yes	No
Inversions		Yes	Yes	No
Complex		Yes	Yes	No
Precision of breakpoints		Single base	A short ambiguous range	Single base
Genotypes of SV events		Yes	No	Yes
Simulated data				
False-positive rate		1.20%	9.1–10.3%	<2%
False-negative rate		9.60%	26–32%	~20%
Inaccessible missing rate[b]		24.10%		
Experimental validation				
False-positive rate		5.60%	11–22%	Not evaluated
NA18507 genome				
No. of calls overlapping[c] with those in previous studies[17,19]/no. of total SV calls	Insertions	52,739/87,457 (60.2%)	5,336/19,305 (27.6%)	124,559/142,908 (87.2%)
	Deletions	42,222/56,074 (75.3%)	4,970/27,092 (18.3%)	133,974/146,843 (91.7%)
YH genome				
No. of calls overlapping[c] with assembly-based calls/no. of total SV calls	Insertions	80,719	28,935/434,984 (6.7%)	21/1,030 (2.0%)
	Deletions	51,711	2,130/177,282 (1.2%)	15/356 (4.2%)
	Complex & Inversion	743	0/247	0/0

Except for structural variation (SV) calls for the YH genome, metrics for BreakDancer and pIndel were reported in their original publications.

[a]Based on simulations reported in initial publication. [b]Defined as SVs in regions that could not be assembled (for Asm), with <2 anomalous mapped read pairs (for BreakDancer) or with <4 gapped aligned reads (for pIndel). [c]Minimum 50% reciprocal overlap between SV calls.

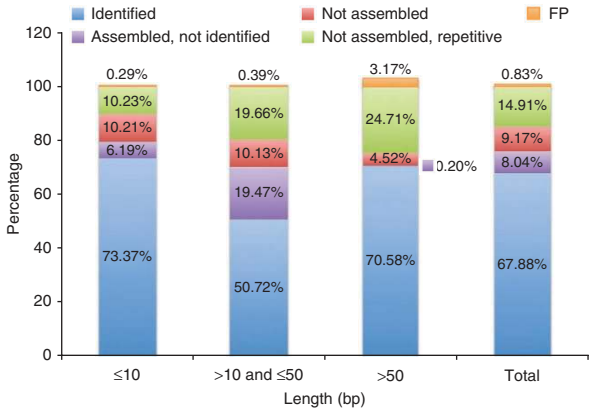

Figure 2 Simulation details. Structural variations were clustered into three parts to evaluate the sensitivity of different length ranges. Identified structural variations (blue) are comparatively higher than those structural variations with sequence assembled but not identified (violet). Most of the missing structural variations are due to the loss of the sequences (not assembled, red) and >10% of them contain repetitive sequences (green). The false-positive rate (orange) is very low for short (≤10) but for intermediate structural variations (>10 and ≤50) and long (>50) structural variations, it is higher due to increased complexity in gapped alignment and statistical analysis respectively.

75.3% of the insertions and deletions, respectively, had been reported previously[17,19] (**Table 2**). Among the indels identified in a previous study[19] but missing in our call set, 64.1% overlap with highly repetitive sequences. BreakDancer identified fewer structural variations of which a smaller percentage agreed with the previous studies. This is likely because BreakDancer excludes indels <10 bp in length owing to limitations of the algorithm, yet most indels are <10 bp. pIndel[42] identified more indels than our method, a greater percentage of which agreed with previous findings. However, pIndel cannot identify large insertions, inversions and complex rearrangements. In contrast, our method identifies these types of variants, including balanced variants[43].

We also compared the results of applying our method, BreakDancer[20] and pIndel[42] to call structural variations in the YH genome (**Table 2**). For indels, most of the variants identified by our approach, of which only 5.6% are expected to be false positives, were not identified by BreakDancer or pIndel (**Supplementary Notes** and **Supplementary Tables 3** and **4** for additional comparisons with other methods, including array comparative genomic hybridization, assemblies obtained using Sanger-sequencing data and a hybrid strategy[23] that incorporates the results of several approaches). These discrepancies between our method and BreakDancer and/or pIndel in the structural variation predictions could be attributed to the varied read lengths (30–75 bp) and library insert sizes[44] used to sequence the YH genome. In summary, our approach provides an accurate method to determine structural variations of different lengths and types. The method is complementary to, and in some cases outperforms, existing methods.

Genome-wide sequence patterns of structural variations

Establishing an accurate and less-biased structural variation map allowed us to investigate genome-wide sequence patterns of structural variations.

We first examined the distribution of structural variations in the YH and NA18507 genomes (**Supplementary Fig. 4**) and found that in both genomes heterochromatic regions (defined as centromeres and telomeres annotated by UCSC hg18) had a higher density of variations. By counting the numbers of shorter (≤10 bp) and longer structural variations (>10 bp) in 1 Mb sliding windows across both genomes, we identified 386 and 330 regions, corresponding to 340 and 299 Mb of sequence, respectively, that had significantly different numbers of structural variations (Fisher's exact test, $P < 0.01$) between the two genomes.

We found 244 intra- and 26 interchromosome transpositions in the YH genome relative to the NCBI36 reference, and 217 intra- and 10 interchromosome transpositions in the NA18507 genome. Notably, 87.4% of the transposition units in the YH and 84.6% in the NA18507 genome did not contain known transposable elements, such as SINEs (short interspersed transposable elements), LINEs (long interspersed transposable elements) or LTRs (long terminal repeats). This finding is consistent with previous observations that most known transposable elements cannot be assembled by short-read *de novo* assemblers; however, it does indicate that nontransposable element transposition could be a notable part of transposition events. We did not observe any significant consensus motifs in the nontransposable element transposition sequences. These types of transpositions may arise from duplication and recombination events[45]. Alternatively, transposons might have been involved but were then mobilized to other parts of the genome. Other than the structural variation peak due to *Alu* insertions or deletions (**Fig. 1b**), we did not observe any significant patterns, such as chromatin-associated periodicities, of insertions and deletions related to nucleosomes structures[46].

We also analyzed the frequency of observing structural variations in protein-coding genes (http://www.uniprot.org/downloads) and *Alu* elements (**Fig. 3**). As expected, we found fewer structural variations in genes as compared with the whole genome, whereas transposons occurred at a higher rate. Of note, the structural variation rate was not evenly distributed across the different functional structures of genes (**Fig. 3a**). Untranslated regions (UTRs) showed a higher structural

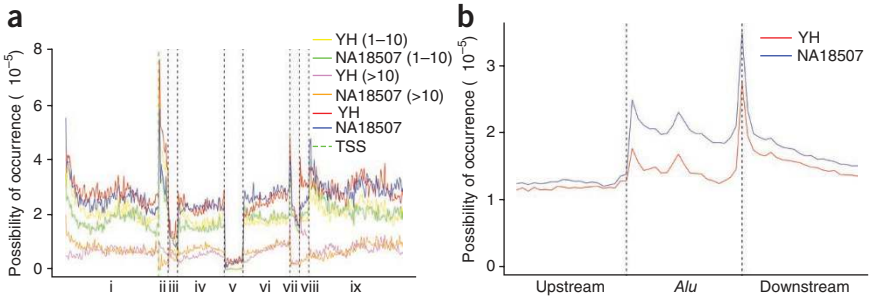

Figure 3 Canonical structural variation profiles of genes and *Alu* elements in YH (red) and NA18507 (blue) genomes. (**a**) The canonical gene structure is defined by nine different features, denoted by the following on the *x* axis: i, upstream; ii, 5′ UTR; iii, first exon; iv, first intron; v, internal exon; vi, internal intron; vii, last exon; viii, 3′ UTR; ix, downstream. *Y*-axis represents the possibility of occurrence of structural variation event per base. Each feature of various length of coding genes was analyzed separately and fitted into equal numbers of bins. Each dot in the respective lines denotes the moving average of 5 bins. Structural variations are classified as 1–10 bp (YH, yellow; NA18507, green) and >10 bp (YH, violet; NA18507, orange). TSS (green dashed line), transcript start site. (**b**) *Alu* transposons with 2 kbp upstream and downstream region. The total probability of structural variation occurrence within *Alu* element is higher than upstream and downstream for both YH (red) and NA18507 (blue).

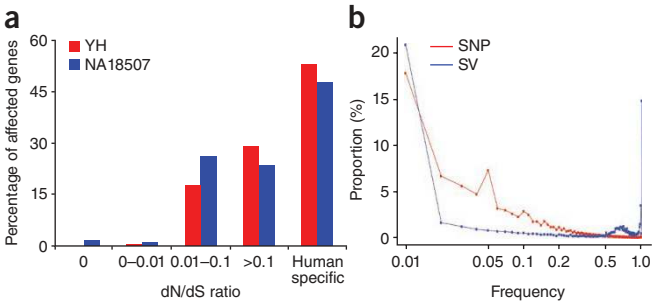

Figure 4 Selection pattern of structural variations. (**a**) Conservation level of structural variation-containing genes of YH (red) and NA18507 (blue) genome. Structural variation-containing genes were categorized by dN/dS ratio according to a comparison between the gene sets of human and mouse genomes from UCSC browser. Two sets were aligned by BLAST. Results with e-value < 1e-20 and identity >90 were included. To avoid double counting, the best results were selected from every aligned region for synonymous and nonsynonymous mutation detection. (**b**) A comparison between the frequency spectrums of identified structural variations and published 1000 Genomes SNP set revealed the excess of very low frequency structural variations. A higher proportion of structural variation (blue) than SNPs (red) is observed at very low frequency.

variation rate than previously reported[47], and the coding sequences had a lower structural variation rate than introns. Notably, *Alu* sequences had more structural variation events than their flanking intergenic regions, and had a spike at each end. This agrees with our expectation that *Alu* sequences undergo transposition more often as intact elements (**Fig. 3b**). In contrast to all the above genomic features, in the 531 annotated micro (mi)RNA regions (http://www.ncbi.nlm.nih.gov) in the human reference genome NCBI36, we identified one structural variation in the YH and three structural variation events in the NA18507 genome. As the functions of miRNAs are sensitive to their length and motifs, these structural variations could have potential deleterious impacts. However, because of the small sample size, more individual genomes and more comprehensive miRNA databases are required to draw solid conclusions.

Annotation of structural variations

We considered the potential functional effects of structural variations. First, we checked for overlaps with repetitive sequences of different lengths (**Supplementary Fig. 5**). Structural variations of 200–400 bp showed the greatest overlap with repetitive sequences (86.5% and 87.6%, respectively, for YH and NA18507), in concordance with previous findings[37]. Next, we annotated the protein coding genes that overlapped with structural variations as they may have substantial functional consequences owing to the potential dramatic effect of the variations on gene structure. We found 8,784 (NA18507) and 8,642 (YH) genes in the two genomes contain structural variations in their gene body. Of these genes, 233 (NA18507) and 281 (YH) had structural variation in the exon sequences (**Supplementary Fig. 6**). To evaluate the potentially deleterious consequences of structural variations in genes, we checked the conservation level (dN/dS ratio in a comparison of the human and mouse genomes) of these structural variation–containing genes (**Fig. 4a**). As expected, more conserved genes showed fewer structural variation events. However, there were 42 and 59 strongly conserved genes (dN/dS ≤ 0.1) that contained structural variations in the YH and NA18507 genomes, respectively. Gene ontology (GO) classifications of these genes showed they belong to ubiquitin, zinc ion binding and nucleus GO categories (**Supplementary Table 5**). Of further interest, many (47.8%) of the genes whose exons contained a structural variation that could be identified using our *de novo* assembly-based approach were missing in dbSNP v.130 (http://www.ncbi.nlm.nih.gov), indicating the usefulness of this method for gaining a more comprehensive view of structural variations in the human genome.

Population distribution of structural variations

Data released as part of the 1000 Genomes Project pilot study[31] provided us with an opportunity to assess the population profile of structural variations detected from the YH and NA18507 genomes. In total, 106 individuals were available for profiling, including 20 Yoruba from Ibadan, Nigeria (YRI), 33 of European ancestry in Utah (CEU), 40 Han Chinese individuals in Beijing (CHB) and 13 Japanese individuals in Tokyo (JPT). A comparison of the frequency spectrum of these structural variations to the published 1000 Genomes SNP set showed an excess of low frequency structural variations[31] (**Fig. 4b**). Similar excesses were observed when comparing nonsynonymous and synonymous substitutions[48]. This result suggests that structural variations are more specific to individuals than are SNPs in humans. Furthermore, structural variations in coding sequences showed a negative correlation between their potential deleterious effects (which are positively correlated with the length of structural variation) and the frequency among the three populations (**Supplementary Fig. 7**). These results indicate that structural variations tend to be under stronger negative selection than are SNPs.

DISCUSSION

We have demonstrated the feasibility and power of identifying structural variations in the human genome by gapped alignment of whole-genome shotgun *de novo* assemblies to a reference genome. We devised a metric, the S/P ratio, to reduce the false-positive rate. By solving complex rearrangements and defining breakpoints of structural variations, we were able to provide a relatively unbiased map for this more refractory type of genetic variation in two human genomes. This capability should facilitate the study of structural variations and their influence on genome evolution and biology. A large portion of the structural variations and patterns identified here, especially those that occur in genes, have not been detected in previous studies using the same genomes. This supports the need to assess and study structural variations using a whole-genome assembly method.

Computational simulations and experimental validation suggest that our results are accurate and that *de novo* assembly can identify structural variations of a wider range of lengths in comparison with previous methods. The structural variation maps of the human genomes have enabled us to initially characterize the genomic patterns of structural variations and their relationship with a variety of genomic features. Many of the observations, for example, genome-wide distribution or canonical analyses of genes and transposons, agree with our prior expectation based on the potential functional impact of structural variations.

Our structural variation maps are still incomplete in several ways. First, we did not focus on heterozygous structural variations in this study because currently available whole-genome *de novo* assembly methods cannot reconstruct haplotypes. Notably, disease-causing structural variations are more likely to be heterozygous (e.g., somatic mutations in cancer samples). Algorithms capable of assembling diploid or even polyploid genomes would extend the applicability of our approach. Second, we were unable to identify variants in highly repetitive sequences, as shown in both simulated and real data, largely because these regions could not be assembled. At present, it is advisable

to apply a combination of different approaches to most comprehensively identify structural variations. But we expect that our alignment and filtering strategies for structural variation identification should be directly applicable to better assemblies that result from improved sequencing technologies and assembly algorithms, potentially proving to be the optimal method to determine structural variation.

Our observation suggests that structural variations are more specific to individuals than SNPs are. Thus, defining structural variations will be of considerable importance for future analyses of personal genomes, as structural variations may underlie phenotypic differences between individuals. Our results suggest that the design of future medical genomics studies and the realization of 'personalized medicine' will require consideration of all the different kinds of genetic variations and their effects on disease and other phenotypes. The reduced bias of our approach and its ability to precisely resolve structural variation breakpoints, in comparison with mapping-based resequencing methods, highlight the need to assemble *de novo* many more human genomes in the future.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology.

Accession codes. DDBJ/EMBL/GenBank: ADDF000000000 (YH) and DAAB000000000 (NA18507). The versions described in this paper are the first versions, ADDF010000000 (YH) and DAAB010000000 (NA18507). NCBI: sequencing reads of YH genome, NCBI Short Read Archive SRA009271. The assembled genomes and all of the associated analyses are freely available at http://yh.genomics.org.cn/.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

This work was supported by a 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. This project is also funded by the Shenzhen Municipal Government and the Local Government of Yantian District of Shenzhen. The 1000 Genomes Project Consortium provided the data for population analysis. AIFB is supported by Diabetes UK, the Wellcome Trust, the Medical Research Council and the Comprehensive Biomedical Research Centre, Imperial College Healthcare NHS Trust. Thanks to X. Wang from School of Biosciences & Bioengineering, SCUT, for his excellent coordination. Thanks to J. El-Sayed Moustafa for her help analyzing the experimental validation data. L. Goodman, S. Edmunds and A. Basford edited the manuscript.

AUTHOR CONTRIBUTIONS

Jun W., Jian W. and H.Y. managed the project. Jun W., Y.L., R. Luo designed the analyses. Y.L., R. Luo, R. Li, H. Zheng, H. Zhu, H.W., H.C., B.W., S.H., H.S., F.Z., H.M., S.F., A.J.d.S., A.I.F.B., W.Z., H.D., L.J.M.C., S.L., L.B. and K.K. performed the data analyses. G.T., J.L. and X.Z. performed the sequencing. Jun W., Y.L. and R. Luo wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/nbt/index.html.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
3. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
4. Hinds, D.A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079 (2005).
5. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nat. Genet.* **37**, 129–137 (2005).
6. Ben-Shachar, S. *et al.* 22q11.2 distal deletion: a recurrent genomic disorder distinct from DiGeorge syndrome and velocardiofacial syndrome. *Am. J. Hum. Genet.* **82**, 214–221 (2008).
7. Futreal, P.A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
8. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
9. Mitelman, F., Johansson, B. & Mertens, F. The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer* **7**, 233–245 (2007).
10. Frazer, K.A., Murray, S.S., Schork, N.J. & Topol, E.J. Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* **10**, 241–251 (2009).
11. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
12. Chanock, S. High marks for GWAS. *Nat. Genet.* **41**, 765–766 (2009).
13. Hirschhorn, J.N. & Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**, 95–108 (2005).
14. Campbell, P.J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40**, 722–729 (2008).
15. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
16. Korbel, J.O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
17. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
18. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
19. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
20. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6**, 677–681 (2009).
21. Lam, H.Y. *et al.* Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat. Biotechnol.* **28**, 47–55 (2010).
22. Conrad, D.F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
23. Pang, A.W. *et al.* Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.* **11**, R52 (2010).
24. 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).
25. Wong, K., Keane, T.M., Stalker, J. & Adams, D.J. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biol.* **11**, R128 (2010).
26. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
27. Simpson, J.T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123 (2009).
28. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
29. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
30. Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. USA* **108**, 1513–1518 (2010).
31. Consortium, T.G. A map of human genome variation from population scale sequencing. *Nature* **467**, 1061–1073 (2010).
32. Harris, R.S. *Improved pairwise alignment of genomic DNA*. PhD thesis, Penn State Univ. (2007).
33. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
34. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2009).
35. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
36. 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).
37. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
38. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics*, chapter 4, unit 4.10 (Wiley, 2009).
39. Alkan, C., Sajjadian, S. & Eichler, E.E. Limitations of next-generation genome sequence assembly. *Nat. Methods* **8**, 61–65 (2011).
40. Alkan, C., Coe, B.P. & Eichler, E.E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).

41. Kidd, J.M. *et al.* Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nat. Methods* **7**, 365–371 (2010).
42. 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).
43. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
44. Li, R. *et al.* Building the sequence map of the human pan-genome. *Nat. Biotechnol.* **28**, 57–63 (2010).
45. Lam, H.Y. *et al.* Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat. Biotechnol.* **28**, 47–55 (2010).
46. Travers, A.A. & Klug, A. The bending of DNA in nucleosomes and its wider implications. *Phil. Trans. R. Soc. Lond. B* **317**, 537–561 (1987).
47. Chen, F.C., Chen, C.J., Li, W.H. & Chuang, T.J. Human-specific insertions and deletions inferred from mammalian genome sequences. *Genome Res.* **17**, 16–22 (2007).
48. Yi, L. Resequencing of 200 human exomes identifies an excess of low frequency non-synonymous coding variants.pdf. *Nat. Genet.* **42**, 969–972 (2010).

ONLINE METHODS

Public data used. The NCBI human reference genome (NCBI36), RefSeq mRNA, dbSNP v.130 and protein sequences were downloaded from the NCBI database (http://www.ncbi.nlm.nih.gov). HuRef assembly were downloaded from http://huref.jcvi.org/. Protein sequences and annotations were downloaded from the UniProt database (http://www.uniprot.org/downloads). Read sequences of sample NA18507 were provided by Illumina, which is also publicly available in the NCBI Short Read Archive (accession number SRA000271).

Identification of structural variations. We pre-aligned all assembled scaffolds to NCBI36 by BLAT[49] V. 30 with the –fastMap and -maxIntron = 50 options enabled. Each hit indicates a possible alignment between a chromosome and scaffold; scaffolds that pre-aligned to identical chromosomes are grouped as scaffold sets. These sets were aligned to corresponding chromosomes by a modified version of LASTZ[32] based on V1.01.50, with high-scoring segment pairs chaining option, ambiguous 'N' treatment and gap-free extension tolerance up to 50 kbp options enabled. Scaffolds with no hits in pre-alignment were aligned to the whole human reference genome with the same options. Inaccurately predicted gaps in assembly, misalignments and three types of complex alignment errors were corrected. With correct alignments, the best hit of every single location on chromosomes was chosen by the utility "axtBest"[33] based on a dynamic-programming algorithm, with the same substitution matrix adopted in alignment. We then selected the hits contributing most to the co-linearity between a scaffold and a chromosome if two or more alignments overlap at the same locus in a chromosome. To guarantee this, each base pair in the reference should be used no more than once. These best alignment hits with gapped extensions include insertions (gap opened in reference since corresponding genotype exists in scaffold) and deletions (vice versa). Software source code for structural variant detection is available at (**Supplementary Data Set 1**, http://soap.genomics.org.cn or http://yh.genomics.org.cn/download.jsp, item 17).

Validation of structural variations. Identified structural variations are classified according to a combination of two approaches.

For structural variation events ≤50 bp. Gapped alignments of all available reads of YH and NA18507 to NCBI36 were performed by BWA using settings that prohibited any gaps >50 bp. Candidate loci were extracted with flanking 150 bp sequences in NCBI36. In particular, genotypes of insertions were inserted into the extracted sequences first. Therefore, for an insertion to be validated, a candidate locus should be aligned without gaps opened, and for a deletion, gaps should be opened with length and the genotype in concordance with predictions. We required at least four reads to make us confident enough of a candidate locus.

For structural variation events >50 bp. The S/P ratio is defined as the ratio between normally aligned and single-end aligned reads of a single base in pair-wise alignment. Short insert-size paired-end reads of the YH and NA18507 genome were aligned to the NCBI human reference genome and their assembly, respectively, by SOAPaligner[50]. S/P ratios were extracted from the alignment results produced by SOAPcoverage (a utility of SOAP package, available at http://soap.genomics.org.cn). To evaluate the overall S/P ratio of each identified structural variation, we calculated the number of mapped paired-end reads that had the expected orientation and insert size (defined as paired-end reads) and those that had an unexpected orientation and insert size (defined as single-ended reads, but were originally from the paired-end library) in both 50-bp flanking regions of each structural variation. Then we calculated the *P*-value by performing Fisher's exact test to test whether the S/P ratio of each structural variation and the S/P ratio of the whole genome are significantly different. Those structural variations with length over three times the s.d. (note that the s.d. may vary with difference sequencing library construction protocols) of the insert-size (about 30 bp—ten-base-pair s.d. multiplied by three—for 200-bp insert-size library) are classified as validated when (i) the *P*-value < 0.05; and (ii) their depths are concordant with their type. A structural variation with a length under 3 times the s.d. of the insert size would be evaluated only by its depth. The average depths between two breakpoints of a deletion in reference or an insertion in assembly were also calculated. Deletions with average depth under a half of the average depth of the whole reference genome and insertions with average depth over a half of the average depth of the whole assembly would be defined as S/P ratio validated.

Fosmid sequences were also used to validate our structural variation results (Fosmid sequences of YH genome available at http://yh.genomics.org.cn). First, we aligned the Fosmid sequences to the genome we assembled using LASTZ, the results of which indicate the concordance between whole genome *de novo* assembly and local assembly. Local assembly would gain less interference from paralogous sequences, which dominantly produce misassembly in whole genome *de novo* assembly. We then selected those Fosmids that had aligned regions overlapping with identified structural variations. A structural variation would be defined as validated if: (i) the Fosmid (including the structural variation) was linearly aligned with the reference; (ii) the breakpoints of the structural variation defined by both Fosmid and whole-genome assembly were concordant; and (iii) the 50-bp flanking sequences of the structural variation in the Fosmid were >90% identical to those in whole-genome assembly.

Simulation of structural variations. We simulated 5,000 homozygous indels ranging from 1 to 50 kbp at random loci (including regions with repetitive sequences) in HuRef assembly chromosome 17. We chose with equal probability between an insertion or deletion, and the length distribution of simulated indels was determined following observations reported in a previous study on NA18507 (ref. 19). Each base in inserted sequences had an equal possibility to be one of the four bases. SNPs were also introduced at a frequency of 0.1%.

We then generated five sets (total 50× coverage) of simulated paired-end reads with read lengths of 35 bp. Mean insert sizes of the read pairs were taken from typical sets of real paired-end data from the Illumina Genome Analyzer, including (i) 210 bp, s.d. 10 bp (20×); (ii) 517 bp, s.d. 19 bp (10×); (iii) 2,522 bp, s.d. 201 bp (10×); (iv) 6,036 bp, s.d. 230 bp (5×); and (v) 9,588 bp, s.d. 636 bp (5×). For all reads, we took sets of base quality values from a previous study on NA18507 (ref. 19) and introduced random substitution errors into the simulated reads at the rate of assigned base quality values.

Simulated reads are then assembled and analyzed by our approach. False positives are defined as those structural variation events identified and validated in the final structural variation set but not in the list of simulated structural variations. False negatives are defined as those simulated events we could not identify or validate in the final structural variation set. False-negative cases were extracted with 50-bp flanking sequences and masked by RepeatMasker with parameter "-s" enabled. Those cases with over half of the bases masked by RepeatMasker are defined as "contained repetitive elements."

Population profiling of structural variations. Raw reads were downloaded from the 1000 genomes project (ftp://ftp.1000genomes.ebi.ac.uk and ftp://ftp-trace.ncbi.nih.gov/1000genomes/), which then were aligned to the NCBI36 and the YH and NA18507 assembled scaffolds. To be defined as an structural variation in an insertion sequence, we required that at least one read was aligned across the breakpoints of a particular structural variation at the scaffold and that the inserted sequence gained coverage >80%. To be defined as an structural variation in a deletion sequence, we required that at least one read was observed across the breakpoint of structural variation at the scaffold and that the deleted sequences in NCBI36 were aligned with coverage <20%.

Comparative genomic hybridization (CGH) array experiments and analysis. We assayed the YH sample using Agilent Technologies standard 244K CGH arrays, using both a local anonymous female sample and a pooled human DNA reference from Promega (http://www.promega.com/country.aspx?returnurl=http%3A%2F%2Fwww.promega.com%2Fproducts%2Fbiochemicals-and-labware%2Fnucleic-acids%2Fgenomic-dna%2F). A total of three array CGH experiments were carried out:

1. YH sample versus anonymous reference.
2. YH sample versus female Promega sample.
3. Anonymous reference versus female Promega sample.

By comparing the aberration lists generated from the three experiments by Agilent CGH array software "DNA analytics," we determined which copy number variations (CNV) were likely to be in the YH sample (as opposed to

being in one or the other of the reference samples). This gave a list of 144 CNVs, consisting of 42 multi-probe and 102 single-probe aberrations (**Supplementary Data Set 2**, **Supplementary Notes** and **Supplementary Fig. 8**).

49. Kent, W.J. BLAT–the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
50. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009).

doi:10.1038/nbt.1904