BreakDancer: an algorithm for high-resolution mapping of genomic structural variation

Ken Chen[1], John W Wallis[1], Michael D McLellan[1], David E Larson[1], Joelle M Kalicki[1], Craig S Pohl[1], Sean D McGrath[1], Michael C Wendl[1], Qunyuan Zhang[2], Devin P Locke[1], Xiaoqi Shi[1], Robert S Fulton[1], Timothy J Ley[1], Richard K Wilson[1], Li Ding[1] & Elaine R Mardis[1]

Detection and characterization of genomic structural variation are important for understanding the landscape of genetic variation in human populations and in complex diseases such as cancer. Recent studies demonstrate the feasibility of detecting structural variation using next-generation, short-insert, paired-end sequencing reads. However, the utility of these reads is not entirely clear, nor are the analysis methods with which accurate detection can be achieved. The algorithm BreakDancer predicts a wide variety of structural variants including insertion-deletions (indels), inversions and translocations. We examined BreakDancer's performance in simulation, in comparison with other methods and in analyses of a sample from an individual with acute myeloid leukemia and of samples from the 1,000 Genomes trio individuals. BreakDancer sensitively and accurately detected indels ranging from 10 base pairs to 1 megabase pair that are difficult to detect via a single conventional approach.

Genomic structural variation is commonly considered to be any DNA sequence alteration other than a single nucleotide substitution[1]. Instances of structural variants in germ and somatic cells contribute, respectively, to heritable genetic diseases[2,3] and cancers[4–6]. Many types of structural variation exist, including insertion-deletions (indels), copy number variants (CNVs), inversions and translocations. Many inherited CNVs (>30 kb) have been discovered using array comparative genomic hybridization[7] and high-density single-nucleotide polymorphism arrays[8]. Alignment of DNA sequences from different sources has been used to identify small or balanced rearrangements not detectable by arrays[9,10]. Recent sequencing and assembly of individual genomes have revealed larger numbers of structural variants than originally expected, especially in the smaller size range (<1 kilobase (kb))[11,12]. However, precise characterization and genotyping of structural variants are still difficult and expensive because of limitations in sequencing technology and detection methods.

Much of the recent advances in structural variation detection can be attributed to next-generation sequencing instruments[13], which have dramatically economized paired-end, whole-genome sequencing. One widely used instrument, the Illumina Genome Analyzer II, uses 100–500 base pair (bp) DNA fragments and requires little input DNA (~1 μg) for sufficient genome-wide coverage. Recent whole-genome resequencing projects[14,15] have obtained paired-end sequence coverage of ×20–40 and have predicted thousands of structural variants using end sequencing profiling (ESP) methods that discerns variants via perceived anomalies in the separation lengths or orientation of aligned read pairs[16,17].

Many substantive issues regarding the analysis of paired-end data, however, remain unresolved. Open questions include whether the procedures and heuristics established for fosmids and bacterial artificial chromosomes can be extrapolated to short inserts, how the expected false positive and negative rates vary with coverage, insert size and read length, and how prediction confidence should be established. As next-generation sequencing data begin to dominate whole-genome resequencing projects, there is a pressing need both to obtain precise answers and to provide practical solutions for data analysis.

Here we address these questions using a combination of computational and experimental approaches. Our software package, collectively called BreakDancer consists of two complementary algorithms. The first, BreakDancerMax, provides genome-wide detection of five types of structural variants: deletions, insertions, inversions and intrachromosomal and interchromosomal translocations from one or a pool of DNA samples sequenced by Genome Analyzer II (**Fig. 1**). The second, BreakDancerMini, focuses on detecting small indels (typically 10–100 bp) that are not routinely detected by BreakDancerMax. Both algorithms (**Supplementary Software**) support pooled analysis that integrates evidence across multiple samples and libraries. In a family- or a population-based study, pooling enhanced the detection of common variants. In a tumor and normal sample paired study, it improved the specificity of somatic variant prediction through effective elimination of inherited variants. Together, these algorithms sensitively and accurately detected many structural variants, as demonstrated in both simulation and real data analysis[14,18,19].

[1]The Genome Center, [2]Division of Statistical Genomics, Washington University School of Medicine, St. Louis, Missouri, USA. Correspondence should be addressed to K.C. (kchen22@wustl.edu).

RECEIVED 6 APRIL; ACCEPTED 13 JULY; PUBLISHED ONLINE 9 AUGUST 2009; DOI:10.1038/NMETH.1363

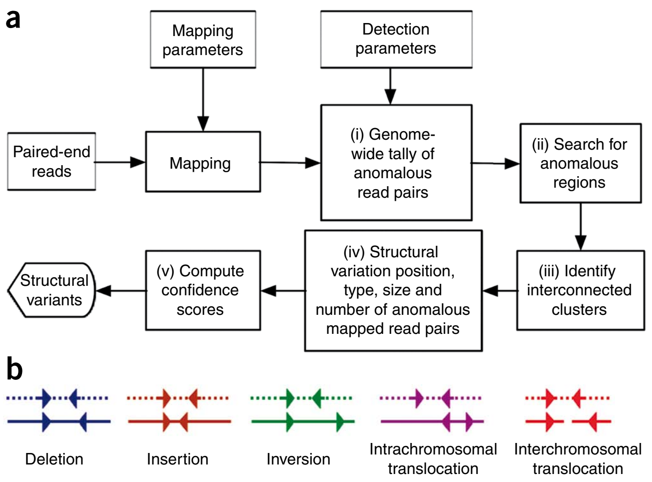
a
Mapping parameters
Detection parameters
Paired-end reads
Mapping
(i) Genome-wide tally of anomalous read pairs
(ii) Search for anomalous regions
(iii) Identify interconnected clusters
(iv) Structural variation position, type, size and number of anomalous mapped read pairs
(v) Compute confidence scores
Structural variants
b
Deletion
Insertion
Inversion
Intrachromosomal translocation
Interchromosomal translocation

Figure 1 | Overview of BreakDancer algorithm. (**a**) The workflow. (**b**) Anomalous read pairs recognized by BreakDancerMax. A pair of arrows represents the location and the orientation of a read pair. A dotted line represents a chromosome in the analyzed genome. A solid line represents a chromosome in the reference genome.

RESULTS

Simulation

To quantify BreakDancer's performance with respect to different parameter settings, we produced synthetic data based on 844 structural variants identified on chromosome 17 of J. Craig Venter's genome[11], which include 425 deletions, 415 insertions and 4 inversions ranging from 20 bp to 7,953 bp. We excluded indels shorter than 20 bp as they are relatively easy to detect using the Smith-Waterman algorithm (**Supplementary Fig. 1**). Many variants in this set occurred in repetitive regions that are difficult to map or assemble (**Supplementary Note**).

We considered a deletion or an inversion as detected if it overlapped 50% reciprocally with a predicted variant. We considered an insertion as detected if its single breakpoint overlapped a predicted variant.

We simulated 50-bp paired-end reads from the chromosome 17 nucleotide sequence of Venter's genome using the software MAQ version 0.7.1 (ref. 20) and obtained 100-fold physical coverage to the US National Center for Biotechnology Information build 36 reference sequence. These reads have normally distributed insert size with a mean size of 200 bp and a standard deviation (s.d.) of 20 bp. We defined anomalously mapped read pairs (ARPs) as those that were confidently mapped by MAQ 0.7.1 (MAQ mapping quality > 10) and had separation distance > 3 s.d. We found that about 365 (43.2%) of the known variants contained 2 or more anomalously mapped reads in their flanking regions and are likely detectable by BreakDancerMax. BreakDancerMax detected 324 (89%) of these 365 variants with a 1.48% false positive rate, including 147 that were shorter than 60 bp (**Fig. 2** and **Supplementary Table 1**).

These 324 variants detected by BreakDancerMax included 214 deletions, 109 insertions and 3 inversions with varying true positive rate in different size ranges and coverages (Online Methods and **Supplementary Fig. 2**). Of the 214 deletions, BreakDancerMax predicted 203 (95%) as deletions with accurate sizes (Pearson's $r = 0.92$) (**Supplementary Fig. 3a**). Of the 109 insertions, BreakDancerMax predicted 72 (66%) as insertions with less accurate sizes ($r = 0.65$) and breakpoints (**Supplementary Fig. 3a,b**). Longer deletions were more accurately predicted in terms of both size and breakpoint.

The confidence score we derived to prioritize BreakDancerMax predictions (Online Methods) demonstrated improved statistical properties when compared to simply using the number of anomalously mapped read pairs, which remains the *de facto* standard metric[21–23]. It provides finer distinction among variants that are supported by identical number of anomalously mapped read pairs (**Supplementary Fig. 4**). It also reduces the result's dependency on the separation threshold and leads to relatively consistent true and false positive rates. (**Supplementary Fig. 5**).

We ran BreakDancerMini on the same data and required the anomalous regions having two-sample Kolmogorov-Smirnov test statistics $D_{nn'} \geq 2.3$, where n and n' are the number of normally mapped reads in the test region and in the whole genome, respectively (Online Methods and **Supplementary Fig. 6**). We observed dramatic improvement in detecting small indels (**Fig. 2**). At 100-fold physical coverage, BreakDancerMini detected 543 (64.3%) variants with a 7.3% false positive rate, including 407 (75.0%) that were shorter than 60 bp. We merged the indels (<100 bp) detected by BreakDancerMini with those detected by BreakDancerMax and obtained a nonredundant set of 683 variants, including 365 deletions, 290 insertions and 21 inversions. Altogether, we detected 621 (74%) of the known variants with a 9.1% false positive rate.

We repeated this simulation under identical conditions but included 10–20-bp indels. In this set, BreakDancerMax alone only detected 24% of the 1,897 known variants with a 7% false positive rate. However, in combination with BreakDancerMini, we detected 68.0% of known variants with a 10.3% false positive rate, 62.6% of which were 10–20 bp. The size of indels appeared to be reasonably accurately predicted throughout the entire range of detection (**Supplementary Fig. 7**).

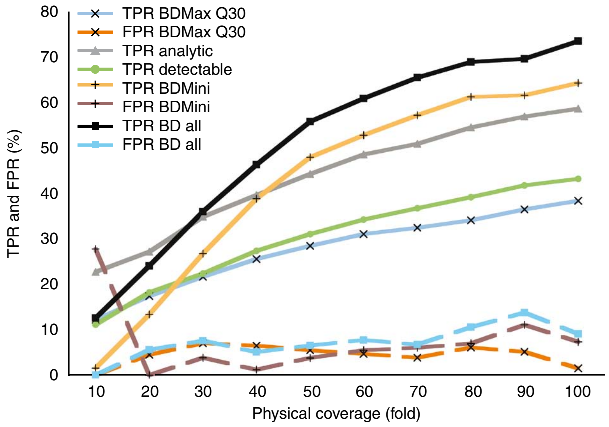

Figure 2 | Performance of BreakDancer in simulation. True positive rate (TPR) and false positive rate (FPR) of BreakDancerMax (BDMax) at the confidence threshold of $Q \geq 30$ (Q30) were analyzed. 'TPR analytic' refers to the percent of variants that can hypothetically be detected by BDMax under an analytic model (Online Methods). 'TPR detectable' is the percent of variants whose flanking regions (300 bp both to the left and to the right) contain 2 or more confidently mapped anomalously mapped read pairs in the MAQ alignment. The performance of BreakDancerMini (BDMini) was characterized by its TPR and FPR. The combined performance (BD all) was obtained by merging the results of these two programs.

Table 1 | Comparison of BreakDancer with other tools

Type	Deletion					Insertion			Inversion
Method	ESP	DIP		Assembly	ESP	DIP		Assembly	ESP
From	Ref. 21	Ref. 21	dbSNP v129	BGI	Ref. 14	Ref. 21	dbSNP v129	BGI	Ref. 21
Size filtering			≥10 bp	≥10 bp			≥10 bp	≥10 bp	
Reported	92	116,395	82,956	107,760	5,704	107,458	82,956	41,134	13
Criteria	Strict	1 bp	1 bp	1 bp	1 bp	1 bp	1 bp	1 bp	1 bp
BreakDancerMax	55/9,202	955/9,202	2,039/9,202	3,123/9,202	5,015/9,202	339/4,901	903/4,901	827/4,901	2/665
BreakDancerMini	21/21,433	4,528/21,433	7,379/21,433	9,344/21,433	1,598/21,433	2,876/17,029	5,083/17,029	3,878/17,029	NA
BreakDancer merged	59/27,092	4,970/27,092	7,998/27,092	10,792/27,092	5,064/27,092	2,983/19,305	5,336/19,305	4,104/19,305	2/655
MPSV weighted	57/8,959	711/8,959	1,332/8,959	2,246/8,959	4,819/8,959	121/5,575	192/5,575	192/5,575	2/504
MPSV unweighted	55/7,599	588/7,599	1,022/7,599	1,835/7,599	4,537/7,599	70/3,772	88/3,772	93/3,772	4/433
Probabilistic	58/8,537	703/8,537	1,217/8,537	2,061/8,537	4,703/8,537	100/7,142	124/7,142	131/7,142	1/181
MoDIL	20/13,147	622/13,147	967/13,147	1,162/13,147	540/13,147	282/3,981	687/3,981	571/3,981	NA

Structural variants predicted by BreakDancer on the Yoruban (NA18507) sample were compared to sets of variants discovered by alternative approaches[14,21]. ESP, large structural variants that were found by analyzing discordant fosmid clone-end alignment; DIP, small deletion or insertion polymorphisms found as gaps in the paired alignment between the fosmid end sequences and the reference; maximum parsimony structural variation (MPSV) weighted, MPSV unweighted and probabilistic refer to three separate sets of structural variants predicted by VariationHunter[24]; MoDIL, the set of variants predicted by MoDIL[25]. Call sets for these tools were downloaded from http://compbio.cs.sfu.ca/strvar.htm and http://compbio.cs.toronto.edu/modil/. The dbSNP version 129 (v129) set refers to indels that are 10 bp or longer in dbSNP version 129. The BGI set refers to 10 bp or longer intracontig indels produced by whole genome *de novo* assembly on the same sample. The 'strict' criteria require the length of the intersection between the validated and the predicted variants to overlap at least 50% of the length of the union of the intervals or the predicted variants to be entirely encompassed by the fosmid interval. Before the slash (/) are the numbers of overlapping variants, after are the numbers of predictions in the corresponding category. NA, not applicable.

Comparison with other methods

We compared BreakDancer with two recently published structural variant detection tools VariationHunter[24] and MoDIL[25]. Notably, these tools both use a different mapping algorithm, MrFast (http://mrfast.sourceforge.net/) than BreakDancer. MoDIL and BreakDancerMini both use the Kolmogorov-Smirnov test[26], but differ in many algorithmic details.

We ran BreakDancerMax and BreakDancerMini on the obtained MAQ map files of the Yoruban genome[14] (Online Methods). We used a separation threshold of 4 s.d. for BreakDancerMax and a threshold of $D_{nn}' \geq 2.3$. for BreakDancerMini. We also required MAQ mapping quality > 10 for both algorithms. BreakDancerMax returned 9,202 deletions, 4,901 insertions and 665 inversions whereas BreakDancerMini returned 21,433 deletions and 17,029 insertions that were shorter than 100 bp. After merging these two sets by position, we obtained a nonredundant set of 27,092 deletions, 19,305 insertions and 665 inversions.

We examined the overlap between the predicted variants with those obtained through alternative approaches (**Table 1**). Altogether, BreakDancer detected 59/92 (64.1%) large fosmid deletions[21], which is comparable to VariationHunter under identical conditions[24]. Among the deletions predicted by BreakDancerMini, 21.1% overlapped at least 1 bp with 4,528 known deletion polymorphisms[21], 34.4% with data from the single nucleotide polymorphism database (dbSNP) version 129 and 43.6% with the intra-contig deletions produced by Beijing Genome Institute (BGI) through whole-genome *de novo* assembly (unpublished data). Among the insertions predicted by BreakDancerMini, 16.9% overlapped with 2,876 known insertion polymorphisms[21], 29.8% with data from dbSNP version 129 and 22.8% with BGI insertions. In general, BreakDancerMini demonstrated substantially higher sensitivity and specificity than VariationHunter or MoDIL. The variant sizes estimated by BreakDancerMini were

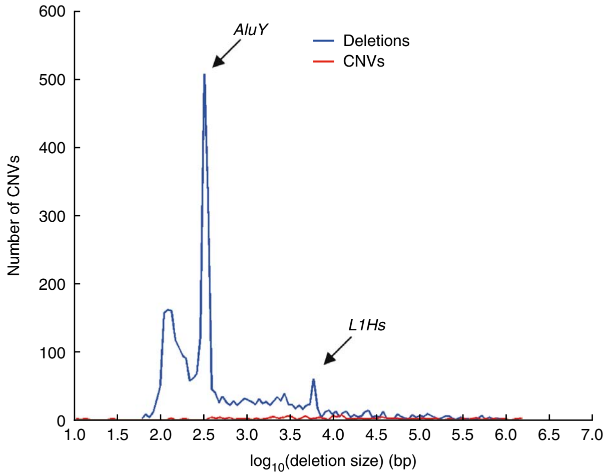

Figure 3 | Size distribution of deletions detected in the genome of an individual with AML. Plotted are size distribution of 3,170 deletions detected by BreakDancerMax from the genome of an individual with AML of 21-fold haploid coverage (deletions) and of 116 inherited CNVs detected using Affymetrix 6.0 array on this sample (CNVs). The deletions range from 58 bp to 959,498 bp. Two signature peaks at 300 bp and at 6,000 bp correspond, respectively, to the *AluY* and the *L1Hs* retrotransposons.

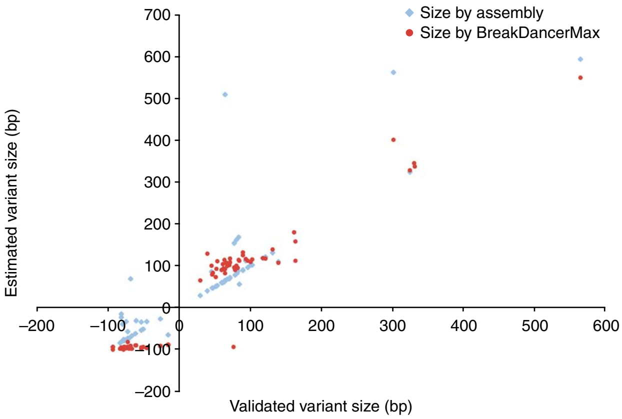

Figure 4 | Accuracy of predicted variant sizes. Plotted are variant sizes predicted by BreakDancer and by local assembly (estimated) versus true sizes determined from the PCR resequencing (validated). Positive sizes represent deletions and negatives represent insertions.

highly correlated with the overlapping deletion and insertion polymorphisms (DIP)[21] ($r > 0.8$).

In addition, 54.3% of the deletions predicted by BreakDancerMax overlapped with 87.7% of the deletions originally reported[14]. Both percentages are higher in comparison to those obtained by VariationHunter[24], possibly because BreakDancerMax uses algorithms similar to those in the original article[14].

Detecting variants in a sample from an individual with AML

We detected variants using data obtained from the tumor and the normal samples of an individual with cytogenetically normal acute myeloid leukemia (AML)[19]. We obtained 21-fold paired-end haploid coverage for both the tumor and the normal genomes, corresponding to 63.5-fold and 39.9-fold physical coverage, respectively.

We jointly analyzed data from six libraries using BreakDancerMax with library-specific separation thresholds and MAQ mapping quality > 35. At a confidence score threshold of $Q \geq 60$, BreakDancerMax detected 7,087 variants, including 3,170 deletions, 1,570 insertions, 1,382 inversions and 965 intrachromosomal translocations (**Supplementary Table 2**). Of these deletions (**Fig. 3**), 46.4% overlapped (50% interval) with known inherited CNVs in the database of genomic variants version 5 (DGV). The percent of overlap became 5–8% higher when culling variants based on the confidence scores instead of the number of anomalously mapped read pairs alone (**Supplementary Fig. 8**). A recent study using Affymetrix 6.0 array had identified 116 inherited CNVs in the genome of the same individual[27], 37 (31.90%) of which overlap with our predictions. These overlapping CNVs range from 131 bp to 1.5 Mbp, with no noticeable bias in size.

We extracted variants that we detected only in the tumor sample and derived 223 putative somatic variants including 100 deletions, 67 insertions (<100 bp), 22 inversions and 34 intrachromosomal translocations. We attempted a local assembly for each of the 167 indels, using the reads mapped to the predicted variant interval (Online Methods). We called variants from the assemblies in 153 of the 167 instances, with 100 confirming the variants (79 both in the tumor and in the normal samples, 17 only in the tumor sample and 4 only in the normal sample).

We experimentally validated these 167 indels (Online Methods). We validated 110 indels (69 deletions and 41 insertions) both in the tumor and in the normal samples, 31 indels were not validated either in the tumor or in the normal samples, and 26 indels were not called owing to low data quality (**Supplementary Table 3**). This suggested a 78% validation rate, excluding the 'no-calls'. Notably, 16 of the 20 deletions that we did not validate had confidence scores below 80 (**Supplementary Fig. 9**). Therefore, the validation rate was 89% at $Q \geq 80$. The size of the deletions determined by BreakDancerMax had good correlation with those determined independently from the validation experiment ($r = 0.867$).

Local assembly improved overall accuracy in that we correctly identified 79 variants in both the tumor and the normal samples. Although the false negative rate of the assembly calls was relatively high (we validated 26 (49%) of the 53 nonvariant calls), the false positive rate was fairly low (we could not validate only 6 (6%) of the variant calls). This observation suggested using assembly for confirmation rather than as a mechanism to limit false negatives. The assembly also improved the size estimation of small indels (**Fig. 4**).

Among the identified insertions, 3 appeared to be ancient alleles that were closer to the chimp than to the human reference genome. In at least 4 inherited deletions we identified, there were stretches of 10–20-bp (A+T)-rich microhomologous sequences inserted between the deletion breakpoints, likely formed by transposons when they inserted into the genome.

We only obtained high-quality validation data for 13 inversions and 6 intrachromosomal translocations. Of these, we validated 4 inversions and 2 intrachromosomal translocations both in the tumor and in the normal samples (**Supplementary Fig. 10**).

Detecting variants in a 1,000 Genomes dataset

We applied BreakDancerMax to the 1,000 Genomes Project[18] data and compared our deletion calls with those that were previously known via fosmid ESP[21] and array comparative genomic hybridization[28] on chromosome 5 of the CEPH (Centre d'Etudes du Polymorphisme Humain) CEU and the Yoruban (YRI) trio individuals.

For each CEU individual's genome, the 1000 Genomes project provided reads from two paired-end libraries with ~15-fold physical coverage (**Supplementary Table 4**). At the threshold of 4 s.d., mapping quality > 35 and $Q \geq 40$, BreakDancerMax detected 125 deletions in the child (NA12878), 79 (63%) of which overlap DGV data. Around 25–35% of known deletions were present in our calls (**Supplementary Table 5**). This percentage increased to 35–45% after we lowered the mapping quality threshold to 10, and the DGV data concordance dropped to 54%. Reducing the separation distance cutoff from 4 to 3 s.d. increased the total number of $Q \geq 40$ predictions by about 20%, but did not increase detection of known variants. Notably, we detected 40–57% of known variants when we jointly analyzed reads from all three individuals with library-specific separation thresholds. There was a substantial overlap among the predicted deletions of the trio individuals: 88/120 (73%) deletions in the father (NA12891) and 98/133 (74%) in the mother (NA12892) were independently detected in the child (NA12878).

We repeated the same set of analyses using data from the YRI trio individuals. Each individual had reads from two paired-end libraries with about 50-fold to 70-fold physical coverage

(**Supplementary Table 4**). At the threshold of 4 s.d., mapping quality > 35 and $Q \geq 40$, 246 deletions were detected in NA19240, 123 (50%) of which overlapped DGV. Around 50–100% known deletions were present in our calls (**Supplementary Table 6**). We detected no additional known variants after lowering mapping quality threshold to 10 or by performing pooled analysis. There was substantial overlap among the deletions of the trio individuals: 168/235 (72%) deletions in the father (NA19239) and 126/164 (77%) in the mother (NA19238) were also independently detected in the child (NA19240).

In contrast to these substantial familial overlaps, the extent of overlap between individuals in different families was lower (31–37%).

DISCUSSION

It is possible to improve BreakDancer's performance by systematically integrating more information in confidence scoring. For example, it may be beneficial to incorporate the mapping quality rather than applying a fixed threshold. Moreover, there is evidence suggesting that integrating read depth may help improve segmentation and genotyping[29], although an effective integration method is yet to be discovered. Our goal is to derive Phred-style quality scores that accurately predict the error probability.

Some types of structural variants, such as inversions and translocations, appeared to be more difficult to detect and validate. Many putative predictions overlapped with regions of tandem or inverted repeat and required further sequence analysis and filtering or the use of additional longer reads and longer inserts. Nonetheless, BreakDancer identified bona fide instances of inversions and intrachromosomal translocations in this study, and somatic inter-chromosomal translocations in our study of samples from individuals with glioblastoma multiforme, ovarian cancer and other subtypes of AML (data not shown).

The algorithms we implemented in BreakDancer are generic and can potentially be expanded to analyze data of different insert sizes or produced by different sequencing technologies. It can also be expanded to analyze paired-end data obtained from mRNA sequencing to identify instances of gene fusion and alternative splicing.

METHODS

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

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

ACKNOWLEDGMENTS

We thank the Genomics of AML Program Project Grant team at Washington University Medical School (US National Cancer Institute P01 CA101937; principal investigagor, T.J.L.) and the 1,000 Genomes Consortium members for providing the data. We thank members of the 1,000 Genomes structural variation group and H. Li for methodology discussions; D. Bentley and M. Ross (Illumina), C. Alkan and J. Kidd (University of Washington), and Y. Li and H. Zheng (Beijing Genome Institute) for providing validation data; and A. Chinwalla, D. Dooling, S. Smith, J. Eldred, C. Harris, L. Cook, V. Magrini, Y. Tang, H. Schmidt, C. Haipek, G. Elliott and R. Abbott for assistance. This work was supported by the National Human Genome Research Institute (HG003079; principal investigator, R.K.W.).

AUTHOR CONTRIBUTIONS

E.R.M., R.K.W., L.D. and T.J.L.: project conception and oversight. K.C.: algorithm design and implementation. J.W.W.: variant assembly. J.M.K., M.D.M. and R.S.F.: experimental validation. C.S.P. and L.D.: primer design. S.D.M. and D.P.L.: Illumina library preparation. Q.Z. and M.C.W.: statistical insight. J.W.W., D.E.L., X.S., and D.P.L.: variant characterization and visualization. K.C., E.R.M., M.C.W., L.D. and J.W.W.: manuscript preparation.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
2. 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).
3. Sharp, A.J. *et al.* A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nat. Genet.* **40**, 322–328 (2008).
4. Futreal, P.A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
5. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
6. Mitelman, F., Johansson, B. & Mertens, F. The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer* **7**, 233–245 (2007).
7. Urban, A.E. *et al.* High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* **103**, 4534–4539 (2006).
8. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
9. Istrail, S. *et al.* Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc. Natl. Acad. Sci. USA* **101**, 1916–1921 (2004).
10. Khaja, R. *et al.* Genome assembly comparison identifies structural variants in the human genome. *Nat. Genet.* **38**, 1413–1418 (2006).
11. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
12. Wheeler, D.A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
13. Mardis, E.R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24**, 133–141 (2008).
14. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
15. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
16. Volik, S. *et al.* End-sequence profiling: sequence-based analysis of aberrant genomes. *Proc. Natl. Acad. Sci. USA* **100**, 7696–7701 (2003).
17. Raphael, B.J., Volik, S., Collins, C. & Pevzner, P.A. Reconstructing tumor genome architectures. *Bioinformatics* **19** Suppl 2, ii162–ii171 (2003).
18. Kaiser, J. DNA sequencing. A plan to capture human diversity in 1000 genomes. *Science* **319**, 395 (2008).
19. Mardis, E.R. *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.* (in the press).
20. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
21. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
22. Korbel, J.O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
23. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
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. Lee, S., Hormozdiari, F., Alkan, C. & Brudno, M. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat. Methods* **6**, 473–474 (2009).
26. Stuart, A., Ord, K. & Arnold, S. Tests of fit. in *Kendall's Advanced Theory of Statistics* Vol. 2A 25.37–25.43 (Arnold, London, 1999).
27. Walter, M.J. *et al.* Acquired subcytogenetic deletions and amplifications in adult acute myeloid leukemia genomes. *Proc. Natl. Acad. Sci. USA* (in the press).
28. McCarroll, S.A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40**, 1166–1174 (2008).
29. Chiang, D.Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods* **6**, 99–103 (2009).

ONLINE METHODS

BreakDancerMax. BreakDancerMax starts with the map files produced by MAQ[20] (**Fig. 1a**). Read pairs mapped to a reference genome with sufficient mapping quality are independently classified into six types (**Fig. 1b**): normal, deletion, insertion, inversion, intrachromosomal translocation and interchromosomal translocation. This classification process is based on (i) the separation distance and alignment orientation between the paired reads, (ii) the user-specified threshold and (iii) the empirical insert size distribution estimated from the alignment of each library contributing genome coverage. The algorithm then searches for genomic regions that anchor substantially more anomalous read pairs (ARPs) than expected on average. A putative structural variant is derived from the identification of one or more regions that are interconnected by at least two ARPs. A confidence score is estimated for each variant based on a Poisson model that takes into consideration the number of supporting ARPs, the size of the anchoring regions and the coverage of the genome. The dominant type of associated ARPs in a particular region determines the type of structural variant. The start and the end coordinates are defined as the inner boundaries of the constituent regions that are closest to the suspected breakpoints, and the size is estimated by subtracting the mean insert size from the average spanning distance in each library and then averaging across libraries.

Confidence score estimation. It is important to derive confidence scores that quantify the underlying error probabilities of the predicted structural variants. The accuracy of the score depends on many factors, for example, whether the set of reads is an unbiased sampling of the genome and all alleles, whether the reads are mapped to correct locations and whether the amount of observed evidence is sufficient.

One of the primary signals for the presence of a structural variant is the clustering of ARPs. Therefore, it was important to measure the degree of clustering from the perspective of both depth and breadth. We assumed that under the null hypothesis of no variant, the genomic location of one particular type of insert was uniformly distributed[14]. For studies that define more than one insert type, the number of inserts at a particular location forms a mixture Poisson distribution with each mixture component representing one of the insert types. The statistic that summarizes the degree of clustering of a particular insert type is the probability of having more than the observed number of inserts in a given region:

$$P(n_i \geq k_i),$$

where n_i denotes a Poisson random variable with mean equal to λ_i, i the type of the insert, and k_i the number of observed type i inserts. The estimation of λ_i is straightforward based on uniform assumption:

$$\lambda_i = \frac{sN_i}{G},$$

where s represents the cumulative size of the regions that the ARPs anchor to, N_i the total number of type i inserts in the entire dataset and G the length of the reference genome. N_i is counted directly from the data without assuming any form of insert size distribution. To detect indels, one could define three types of inserts: long, medium and short defined by predetermined thresholds. The task of indel detection is to find deletions from regions that contain significantly more long inserts ($P < 0.0001$) and insertions from regions that contain significantly more short inserts ($P < 0.0001$) in above defined Poisson test.

This probabilistic scoring system can conveniently integrate information from multiple libraries from the same or different individuals using Fisher's method[30] assuming that the m libraries are produced independently:

$$\chi^2_{2m} = -2 \sum_{j=1}^{m} \log_e(P_j),$$

where χ^2 denotes a chi-square distribution of $2m$ degree of freedom and P_j the P value obtained from the j^{th} library.

This makes it straightforward to compute a combined P value from variable insert-size libraries, or from one or multiple individuals to fully harness the statistical power of the pooled data. For convenience of representation, we convert the combined P value to Phred scale using:

$$Q = -10 \log_{10}(P).$$

However, this Q score is not necessarily a Phred quality score, although they should have good correlation.

An analytic model of true positive rate (TPR) in simulation. Assuming that all the reads can be confidently mapped and that the ARPs cannot intersect with the variant breakpoint, we can analytically estimate the number of ARPs that a known structural variant may have

$$\bar{n}^d = Ra \frac{\mu - 2l}{\mu} (1 - G(c^d, \mu + \theta^d, \sigma))$$

for deletions and

$$\bar{n}^i = Ra \frac{\mu - 2l - i}{\mu} G(c^i, \mu - \theta^i, \sigma)$$

for insertions, where $G(.)$ represents the insert size distribution function with mean μ and s.d. σ, size of the deletions θ^d, size of the insertions θ^i, threshold that defines the long inserts c^d, threshold that defines the short inserts c^i, read length l, physical coverage R and allele frequency a.

We can compute the TPR in our simulation using this analytic model that summarizes information about the insert size, read length, coverage and the variant size (**Fig. 1**). With a 200 bp insert library (s.d., 20 bp and read length, 50 bp), 493 (58.69%) of 844 known variants (≥20 bp) on the chromosome 17 of J. Craig Venter's genome would possess 2 or more ARPs (≥3 s.d.) at 100-fold physical coverage. This analytic TPR approaches an asymptote at 180-fold, at which all deletions are detected and at 220-fold, at which 307 (74%) of 415 insertions are detected (**Supplementary Table 7**). For a 400-bp insert library (s.d., 40 bp and read length, 50 bp), the analytic TPR approaches an asymptote at 430-fold, at which all deletions are detected and at 470-fold, at which 87.5% of insertions are detected (**Supplementary Table 8**).

We can explicitly characterize the analytic TPR as a function of variant size and coverage based on the Poisson coverage model:

$$r^d = P(n \geq 2 | \lambda = \bar{n}^d)$$

for deletions of size d, and

$$r^i = P(n \geq 2 | \lambda = \bar{n}^i)$$

doi:10.1038/nmeth.1363

for insertions of size *i*, where $P(\cdot)$ represents the Poisson distribution function.

With these formulas, it can be shown that insertions and deletions shorter than 40 bp are difficult to detect using the above 200 bp insert library owing to the 20 bp s.d. Deletions longer than 60 bp took about 30-fold coverage to reach an asymptote and those longer than 100 bp took only 20-fold (**Supplementary Fig. 11a**). Insertions ranging from 60 bp to 80 bp were relatively easier to detect (**Supplementary Fig. 11b**), but those longer than 100 bp could not be detected at all, as their detection was limited by the insert size and read length of DNA fragments.

BreakDancerMini. BreakDancerMini analyzes the normally mapped read pairs that were ignored by BreakDancerMax. A genomic region of size equivalent to the mean insert size is classified as either normal or anomalous based on a sliding window test that examined the difference of the separation distances between read pairs that are mapped within the window versus those in the entire genome. Similar to BreakDancerMax, a putative structural variant could be derived from the anomalous genomic regions that are interconnected by at least two common read pairs. A confidence score is assigned based on the significance value of the sliding window test. The start and the end coordinates are decided as the outer boundaries of the constituent regions, and the size is estimated using the same approach as for BreakDancerMax.

The sliding window test. We applied a sliding window test to identify anomalous regions that contain read pairs significantly ($P < 0.0001$, Kolmogorov-Smirnov test) different from the entire genome. By default, BreakDancerMini using a fixed window size of $w = \mu + 3\sigma - 2l$ bp and a step size of 1 bp, where μ and σ are the mean and the s.d. estimated from the separation distance of normally and confidently (mapping quality > 40) mapped read pairs, and l is the average read length. A two-sample Kolmogorov-Smirnov (KS) test statistic[26]

$$D_{nn'} = \sqrt{nn'/(n+n')}\,\sup_{x}|F_n(x) - F_{n'}(x)|$$

is computed for each window, where $F_n(x)$ and $F_n'(x)$ are the empirical cumulative distribution function (ECDF) estimated from the normal reads in the window and in the entire genome respectively, and n and n' are the number of reads in each set; x is the separation distance from 1 bp to a maximum size (~300 bp); sup denotes the supremum of the set. Obviously, D_{nn}' objectively measures the difference between the two ECDFs in terms of both location and shape. To model alignment orientation, we computed two statistics D^{+}_{nn}' and D_{nn}' per window using reads that are mapped to the plus and the minus strands respectively. A genomic region is classified as anomalous in either the plus or the minus orientation if the corresponding KS statistic exceeds a user-selected threshold. Overlapping anomalous regions in the same orientation are filtered out and only the highest scoring one is kept. For small indels, the anomalous regions that support the same variant are required to be in the opposite orientations. In principle, this approach works with any insert size distribution and does not require any predetermined separation threshold.

Variant calling based on local assembly. A local assembly of the breakpoints within a suspected variant region can confirm the existence of the structural variant, precisely define the breakpoint locations and determine any inserted sequences that may be present. In our AML study, we assembled reads mapped by MAQ to within 500 bp of the predicted variant boundaries, including unaligned reads whose mates mapped within the region using both Velvet[31] and Phrap. We found that using more than one assembly algorithm increased the chance of assembling a structural variant. If the derived contig sequences cumulatively covered over 75% of the region from which the reads were extracted, we aligned the contigs to a region of the human reference sequence containing the structural variant and 1,000 bp of flanking sequence on either side using cross-match. The resulting pairwise alignments were examined for the existence of breakpoints or gaps. A variant was called if there is a gap or if the tumor and the normal contigs contain consistent breakpoint.

Experimental validation. Experimental validation was performed on putative structural variants in the DNA from the tumor sample from an individual with AML (AML tumor) and normal genomes. Primer3 was used in conjunction with internal software to design and select tailed PCR primers for structural variant validation. Efforts were made to avoid designing primers in repetitive regions and to select primers with average G+C content close to 50% and a predicted melting temperature (T_m) of 60 °C. Primers were selected by hand when automated methods indicated a low likelihood of success. For small insertions, small inversions and deletions of most sizes, PCR primers were designed ~100–200 bp outside of the boundaries of the breakpoints defined by BreakDancer. For large inversions and intrachromosomal translocations, primers were designed with the same orientation as, but 10–200 bp upstream of, any variant supporting read pairs. If a structural variant was supported by both forward and reverse read pairs across both breakpoints, four primers were designed and two separate attempts were made to validate the variant by PCR amplification and Sanger sequencing. Structural variants were considered validated if any single resulting read sequence spanned the predicted breakpoints. No primers were designed for complex events, for example, if conserved repeats spanned or flanked both ends of the predicted breakpoints. Genomic DNA from the tumor and a matched normal blood sample were amplified using standard PCR protocols. Putative small insertions, small inversions and deletions of all sizes were amplified using Amplitaq Gold polymerase (Applied Biosystems). Putative large inversions and intrachromosomal translocations were amplified using Accutaq Hotstart polymerase (Sigma Aldrich). All PCR products were evaluated on a 2% agarose gel. Regardless of yield, all products were sequenced in both directions using Big Dye Terminator (Applied Biosystems) reactions and subsequently loaded on an 3730xl capillary sequencer (Applied Biosystems). The resulting traces were assembled to a reference sequence extracted from the region surrounding the predicted variant site on NCBI build 36 with an additional 1 kbp of flanking 3′ and 5′ sequence. All resulting diploid trace data were manually reviewed and those traces showing unambiguous evidence of homozygous or heterozygous SV were classified as either somatic or germline events, or alternatively, labeled as variants if the somatic status could not be determined due to lack of sequence data from the matched normal sample.

The NA18507 data. We downloaded approximately 3.5 billion end sequences (1.7 billion pairs) of length 36 to 41 bp and insert size 200 bp from the US National Center for Biotechnology Information (NCBI) Short Read archive. This constituted about 42-fold sequence and 120-fold physical coverage of the human genome. We mapped all reads from the 200 bp library to the NCBI build 36.1 reference using MAQ 0.7.1 and obtained 37.2-fold haploid coverage after removing the duplicated reads that had identical outer coordinates. Consistent with the previous reports[24], the obtained insert size distribution was approximately normal with a mean of 209 bp and a s.d. of 13 bp.

The AML data. We constructed four Illumina paired-end libraries from the genomic DNA of the primary tumor cells and two libraries from the normal skin cells of an individual with AML. The mean insert sizes range from 95 bp to 268 bp based on the empirical insert size distributions estimated from the alignment (**Supplementary Table 1**). All libraries had unimodal insert size distributions although the normal DNA libraries had a relatively larger s.d. than the tumor libraries (**Supplementary Fig. 12**). Some libraries have distributions clearly diverged from Gaussian and these can be problematic for variant detection methods that assume normality. For both the tumor and the skin genomes, we obtained 21-fold haploid sequence coverage, corresponding to 63.5- and 39.9-fold physical coverage, respectively. Of the paired-end reads obtained, 67% were 50 bp and the rest between 35 bp and 36 bp. All reads were mapped to the NCBI build 36 human reference sequence using MAQ 0.7.1.

System requirements and software availability. BreakDancer is currently written in Perl and is available at http://genome.wustl.edu/tools/cancer-genomics/. It usually takes 3–5 h and between 200 Mb to 500 Mb memory to analyze one human chromosome at around 50-fold sequence redundancy.

30. Fisher, R.A. Combining independent tests of significance. *Am. Stat.* **2**, 30 (1948).
31. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).

doi:10.1038/nmeth.1363