BMC Genomics

BioMed Central
The Open Access Publisher

DB2: a probabilistic approach for accurate detection of tandem duplication breakpoints using paired-end reads

BMC Genomics 2014, **15**:175 doi:10.1186/1471-2164-15-175

Gökhan Yava¿ (gokhan.yavas@case.edu)
Mehmet Koyutürk (koyuturk@eecs.case.edu)
Meetha P Gould (meetha.gould@case.edu)
Sarah McMahon (sarah.mcmahon@case.edu)
Thomas LaFramboise (thomas.laframboise@case.edu)

ISSN 1471-2164

Article type Methodology article

Submission date 12 June 2013

Acceptance date 18 February 2014

Publication date 5 March 2014

Article URL http://www.biomedcentral.com/1471-2164/15/175

DB^2: a probabilistic approach for accurate detection of tandem duplication breakpoints using paired-end reads

Gökhan Yavaş[1]
Email: gokhan.yavas@case.edu

Mehmet Koyutürk[2,4]
Email: koyuturk@eecs.case.edu

Meetha P Gould[3]
Email: meetha.gould@case.edu

Sarah McMahon[3]
Email: sarah.mcmahon@case.edu

Thomas LaFramboise[3,4,5*]
[*] Corresponding author
Email: thomas.laframboise@case.edu

[1] Department of Epidemiology & Biostatistics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106, USA

[2] Department of Electrical Engineering & Computer Science, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106, USA

[3] Department of Genetics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106, USA

[4] Center for Proteomics and Bioinformatics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106, USA

[5] Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic Foundation, 9500 Euclid Avenue, Cleveland, OH 44195, USA

Abstract

Background

With the advent of paired-end high throughput sequencing, it is now possible to identify various types of structural variation on a genome-wide scale. Although many methods have been proposed for structural variation detection, most do not provide precise boundaries for identified variants. In this paper, we propose a new method, **D**istribution **B**ased detection of **D**uplication **B**oundaries (DB^2), for accurate detection of tandem duplication breakpoints, an important class of structural variation, with high precision and recall.

Results

Our computational experiments on simulated data show that DB^2 outperforms state-of-the-art methods in terms of finding breakpoints of tandem duplications, with a higher positive predictive value (precision) in calling the duplications' presence. In particular, DB^2's prediction of tandem duplications is correct 99% of the time even for very noisy data, while narrowing down the space of possible breakpoints within a margin of 15 to 20 bps on the average. Most of the existing methods provide boundaries in ranges that extend to hundreds of bases with lower precision values. Our method is also highly robust to varying properties of the sequencing library and to the sizes of the tandem duplications, as shown by its stable precision, recall and mean boundary mismatch performance. We demonstrate our method's efficacy using both simulated paired-end reads, and those generated from a melanoma sample and two ovarian cancer samples. Newly discovered tandem duplications are validated using PCR and Sanger sequencing.

Conclusions

Our method, DB^2, uses discordantly aligned reads, taking into account the distribution of fragment length to predict tandem duplications along with their breakpoints on a donor genome. The proposed method fine tunes the breakpoint calls by applying a novel probabilistic framework that incorporates the empirical fragment length distribution to score each feasible breakpoint. DB^2 is implemented in Java programming language and is freely available at http://mendel.gene.cwru.edu/laframboiselab/software.php.

Background

Structural variation is a class of genetic variation that includes insertions, inversions, translocations, deletions, and duplications of segments of DNA. Tandem duplications are serially repeated segments of the human genome which may have repeat units several hundred kilobases in size. Many studies have implicated tandem duplications in a variety of diseases. In one such study [1], it was shown that a subset of ovarian cancers share a marked tandem duplication phenotype with triple-negative breast cancers. An internal tandem duplication of the *FLT3* gene (*FLT3*/*ITD*) is recurrent in acute myeloid leukemia (AML) and myelodysplastic syndrome (MDS) with frequencies of 20 and 3-15%, respectively [2,3]. Additionally, 5% to 10% of patients with AML possess the rearrangement of the mixed-lineage leukemia (*MLL*, also known as *ALL1* or *HRX*) gene as the result of a partial tandem duplication (PTD) [4]. Germline tandem duplications have also been associated with human disease. In one recent study [5], it was shown that a patient and his half-sister with extensive polysyndactyly of the hands and feet, and craniofacial abnormalities carried identical 900-kb tandem duplications of the Indian hedgehog (*IHH*) locus. Another study [6] reported a father and daughter, both with a history of compulsive over-eating in childhood, carrying a small tandem duplication within exon 1 of the *SNURF*/*SNRPN* gene on chromosome 15. These studies underscore the need for computational methods for identifying tandem duplications.

Next-generation sequencing (NGS) technology was first used to detect structural variations by Korbel *et al.* [7]. In that study, the paired-end sequences of two samples' genomes were generated and the read pairs with discordant paired-end orientation and mapped distance were used to find basic structural variations. Subsequently, [8] used NGS to discover genome

rearrangements in tumor DNA. The first genome that was wholly sequenced by a NGS platform was presented in [9], which reported several structural variations.

NGS data provides several sources of information from which methods may detect structural variation, including read depth, paired-end orientation, distance between mapped ends, and pairs where one end is "split" mapped or "one-end anchored" (i.e., its mate is not mapped). PEMer [10], BreakDancer [11], VariationHunter [12,13], GASV [14], and GASVPro [15] use the orientation and the mapped distance between the read pairs to detect insertions, deletions, inversions, and/or translocations. CREST [16] is another method that utilizes split mapped reads as well as paired-end read orientation. The problem of finding novel insertions was also addressed using one-end anchored read pairs in another recent study [17]. In addition, EWT [18] and SegSeq [19] were developed for detecting the genomic regions that differ in copy number between individuals using the depth of single reads in sequence data. Currently, the most well-known methods for detecting the tandem duplications (along with other types of variations) using just the paired-end NGS data include SVDetect [20], CNVer [21], SPANNER [22], inGAP-sv [23], BreakDancer [11], GASV [14] and CREST [16].

For methods that use paired-end reads, an important factor is fragment length, since the two sequenced ends of each fragment will be separated by this length. However, the length of each fragment is not known precisely. Although many of the existing methods assume that fragment length is within a certain range for all fragments [12-14], they do not make use of important information contained in the distribution of these lengths when prioritizing among the predicted breakpoints of the structural variations. If the length of each fragment were known, one could use this information to precisely detect the boundaries of duplications. While precise lengths are not generally available, their general distribution can be derived empirically from *concordantly* mapped reads. Here a read pair is said to be concordantly mapped to the reference genome when the end with a lower mapping coordinate is aligned to the forward strand, the end with the higher mapping coordinate is aligned to the reverse strand (i.e., FR read pairs, where F and R refer to forward and reverse strands, respectively) and the distance between the mapped ends is within an expected range.

Motivated by this insight, here we propose a method **D**istribution **B**ased detection of **D**uplication **B**oundaries (DB^2) that characterizes the distribution of fragment length empirically and utilizes this empirical fragment length distribution to predict the breakpoints of the tandem duplications at a very high resolution with high accuracy and low false positive rate. To the best of our knowledge, none of the existing methods developed for detecting any kind of structural variations utilizes this valuable information for predicting the breakpoints of detected variations. Although we focus on tandem duplications in this manuscript, the proposed framework can easily be extended to detect the boundaries of other structural variations as well.

The general framework implemented by DB^2 is summarized in Figure 1 (see Methods for details). Briefly, DB^2 uses the Binary Alignment/Map (BAM) files obtained by mapping the paired-end read sequences to the human reference genome using BWA [24] (or any other alignment tool that can produce BAM files). The resulting BAM files include orientation information as well as the mapping coordinates for each read pair. Concordant read pairs map in the expected FR orientation, and are thought to correspond to regions that do not differ from the reference genome (in structural terms), whereas pairs with an "everted" RF orientation are indicative of tandem duplications [25].

Figure 1 A flowchart summarizing the framework implemented by DB^2. Since the distances between the aligned ends of the concordantly mapped read pairs can be considered as representatives of the real fragment lengths, we first extract the concordant read pairs from the BAM files and obtain the empirical fragment length distribution using them. The everted (RF) read pairs, which are also extracted, are indicative of tandem duplications. We use each of the RF pairs along with the empirical fragment length distribution to represent the feasible breakpoints of the tandem duplication that induced this RF pair. Next, DB^2 clusters the read pairs that may be induced by the same tandem duplication, and hence finds distinct tandem duplications along with their potential breakpoints. It scores each potential breakpoint by utilizing the empirical length distribution and obtains the breakpoint with the highest score as the putative breakpoint of each tandem duplication. After the conflict resolution step eliminates the likely false positives, the final set of tandem duplications are reported to the user.

DB^2 uses the read pairs that are reported to be concordant by the alignment tool to deduce the empirical fragment length distribution, and the RF read pairs for discovering the tandem duplications along with their putative genomic breakpoint coordinates. To identify the tandem duplications, DB^2 adopts the geometric representation of the putative breakpoints of a tandem duplication that induces a discordant read pair, which was first proposed in the design of GASV [14]. Our method then groups the RF read pairs that are likely to be induced by the same tandem duplication and uses the information extracted out of multiple read pairs along with the empirical fragment length distribution to precisely infer the putative breakpoints of the tandem duplications.

As a final step, we resolve the conflicts among the tandem duplications, which are caused by multiple distinct tandem duplications having overlapping boundaries, by applying an algorithm that relies on the maximum parsimony principle. After the most likely false positive tandem duplications are eliminated in this step, the set of conflict-free duplications are reported to the user. As we show via systematic computational experiments in the Results section, incorporation of fragment length distribution greatly improves our method's ability in fine tuning the breakpoints of identified duplications.

Results and discussion

Simulation procedure

For simulation testing, we have implemented an artificial paired-end read generator using the February 2009 assembly (Hg19) of the human reference genome. Our simulator generates paired-end read sequences that are similar to those of the Illumina/Solexa platform (see Materials and Methods section for details). To evaluate the performance of the proposed method, for each experiment, we inserted 1000 tandem duplications whose lengths (in bases) were drawn from a normal distribution, with a default standard deviation of 100 bp and default mean of 10 Kbp, into the reference genome. For the experimental evaluation of our algorithm, we used four criteria; precision, recall, F_1-score and mean breakpoint mismatch. Precision is defined as the fraction of the number of true tandem duplications (true positives) among all tandem duplications identified by our algorithm (true positives and false positives). In order for a predicted (by our method or other methods) tandem duplication to be considered as a true positive, we required at least 50% mutual overlap of the real and the predicted tandem duplications. Recall is defined as the fraction of true positives among all

tandem duplications in the donor genome (true positives and false negatives). F_1-score is a commonly used aggregate metric in information retrieval that considers both precision and recall. It is defined as the harmonic mean of precision and recall. Mean breakpoint mismatch is defined as the average of total distances (in bp) between the predicted and the real start and end positions of the inserted tandem duplications.

Other methods used for comparison

We compared the performance of our algorithm with that of five other software packages designed to detect structural variations from paired-end NGS data: SVDetect [20], CNVer [21], Breakdancer [11], GASV [14] and CREST [16]. Note that the more recent version of GASV, GASVPro, is not included in the compared methods because it does not support the identification of the tandem duplications. Although SPANNER [22] and inGAP-sv [23] are also able to detect tandem duplications, both of these methods were excluded from the experimental evaluation since SPANNER was not publicly available and inGAP-sv was significantly outperformed by the other methods. For all the methods, we aligned the generated read pair sequences with BWA using the default parameters. The default parameters for CNVer, Breakdancer, CREST and GASV were used, whereas the default values of *window_size* and *step_length* parameters had to be slightly modified in SVdetect to obtain the best performance with the simulation data. We set these two parameters to 1000 and 500, respectively.

Several factors can affect any method's ability to detect a tandem duplication: the average depth coverage of the experiment, the base call error rate, characteristics of the tandem duplications in the donor genome (such as the size of the tandem duplications), properties of the read library (including the distribution of the fragment lengths), and read length. For this reason, we tested the algorithms across various values of six parameters as discussed in the following sections.

Effect of base calling error rate on performance

To evaluate the effect of base call errors, we simulated different error rates using our synthetic data generator by changing each base with a probability that is defined with the base call error rate. As shown in Figure 2A, the precision of our method, Breakdancer and GASV is steady at 99-100% for all base calling error rates. On the other hand, the precision of CNVer decreases dramatically as error rate increases whereas CREST first has a decreasing and then increasing precision performance. Somewhat surprisingly, SVDetect has an increasingly better performance as the base calling error increases. We observed that it can reach at most 97% at the highest level of noise induced in our simulations, which is still lower than DB^2's performance. The positive impact of error rate on precision is likely because the alignment tool will drop spurious mappings as error rate goes up.

Figure 2 Performance as a function of error rate. (A) Precision, **(B)** Recall and **(C)** F_1-score performances of the methods at different base calling error rates are presented. Here the average depth coverage is fixed at 40X.

The recall of our method and SVDetect are almost identical (Figure 2B), whereas CNVer, GASV, Breakdancer and CREST have drastically declining performances with increasing error rate. The decrease in the sensitivities of all methods can be explained by the fact that the alignment tool fails to align increasingly noisy RF reads. Thus, as the error rate goes up, the

effective coverage goes down, and the evidence for the duplications gets weaker, which results in fewer predictions and hence fewer true positives. To validate this claim, we computed the mean number of the read pairs supporting each tandem duplication as the base calling error increases (Additional file 1: Figure S1). As shown in this figure, the support for each tandem duplication significantly decreases due to lower effective coverage as we increase the noise in the data. To assess the overall accuracy of the methods, we present the F_1-score performance in Figure 2C. As mentioned before, F_1-score evaluates the precision and recall performance of each method by aggregating them into a single value for each error rate level. As seen in Figure 2C, our method outperforms all the presented methods in terms of F_1-score at each error rate.

As seen in Figure 3, our algorithm outperforms SVDetect and CNVer in terms of finding the breakpoints of the tandem duplications but CREST is able to identify the exact location of the tandem duplication. Although Breakdancer can attain a mean breakpoint mismatch performance similar to that of our method for low error rates, DB^2 outperforms it by maintaining a robust performance even for very high base calling error rates.

Figure 3 Mean breakpoint mismatch at different base calling error rates. Breakpoint mismatch is calculated as the average number of bases between the real and predicted breakpoints. Average depth coverage is fixed at 40X.

Overall, DB^2 provides the best F_1-score, which represents the aggregate of precision and recall, along with a very good mean breakpoint mismatch that is tolerable as the noise in the data increases.

Effect of depth coverage on performance

Breakdancer, GASV and DB^2 outperform the other three methods in terms of precision across a wide range of coverages. As seen in Figure 4A, those methods' precision stabilizes around 99-100%, whereas precision declines with increasing coverage for SVDetect (this is consistent with SVDetect's declining performance with decreasing error rate, since increased coverage also results in more false mappings) and CREST. CNVer has a rather stable performance around 92.5% as a function of depth coverage. On the other hand, recall for DB^2 and SVDetect stabilizes at around 99% as the coverage increases, whereas GASV, CREST, CNVer and Breakdancer peak at 92%, 85% , 90% and 89%, respectively (Figure 4B). In terms of F_1-score, DB^2 performs much better than all the other methods having a stable score around 98.5% whereas our closest competitor, SVDetect, stabilizes at around 95.5% (Figure 4C). This shows our method's ability to maintain very high precision and recall performances with changing depth of coverage levels.

Figure 4 Performance as a function of depth coverage. (A) Precision, **(B)** Recall and **(C)** F_1-score performances at different average depth coverage levels are shown. Here the base call error rate is fixed at 0.01.

For varying levels of coverage, CREST again attains nucleotide-level accuracy with regard to mean breakpoint mismatch for true tandem duplications whereas our algorithm has a slightly lower performance than that of CREST. On the other hand, DB^2 consistently and substantially outperforms CNVer and SVDetect in terms of this metric (Figure 5). Indeed, DB^2 is able to accurately localize breakpoints to within 15 bases or fewer even at low coverage values. This observation suggests that the use of fragment length distribution indeed

improves accuracy in fine-tuning of the breakpoints, as it gives more importance to breakpoints consistent with a higher frequency fragment length (see Methods for details). On the other hand, Breakdancer and GASV slightly perform worse for low coverage levels but then their performances catch up with the performance of DB^2 for higher coverage values.

Figure 5 Mean breakpoint mismatch at different depth coverage levels. The base call error rate is fixed at 0.01.

Varying levels of coverage directly impact the amount of data available to each method. As shown in the above analysis, DB^2 consistently achieves the best F_1-score and recall performance, but has slightly worse mean breakpoint mismatch performance than that of CREST, even when the data availability is low (i.e., lower coverage levels). Considering the CREST's much lower recall and precision performances, DB^2's average mismatch of 15 base pairs when identifying the boundaries of a tandem duplication is quite tolerable.

Effect of duplication size on performance

For this set of experiments, we increased the size of the tandem duplications starting from 2 Kbp up to 10 Kbp in 2 Kbp increments for each experiment setting. Almost all of the methods have a stable performance in terms of all metrics as we increase the size of each duplication inserted into the donor genome (Additional file 2: Figure S2 and Additional file 3: Figure S3). This is an expected result for DB^2, since as long as the fusion point of a tandem duplication is straddled by a read pair, DB^2 will use this information to identify its breakpoints regardless of duplication size.

Effect of changing properties of the read library on performance

There are multiple important factors during the read library preparation phase of any NGS experiment that can affect the performance of a structural variation identification method. These include (but are not limited to) the distribution of the lengths of the fragments, and the read length.

In order to see the effects of these factors, we conducted a series of experiments by changing the values of read length and fragment length mean/standard deviation during the simulation data preparation. With the exception of CREST, we observe no significant effect on any method's Recall, Precision and F_1-score performance (Additional file 4: Figure S4, Additional file 5: Figure S5 and Additional file 6: Figure S6, respectively). CREST performs poorly in terms of recall for a read length of 50 bp, but then improves for larger read lengths (Additional file 6: Figure S6). In contrast, the precision performance of CREST first deteriorates as we enlarge the reads, and then stabilizes around 70%.

Increasing the mean value of the fragment lengths dramatically decreases the mean boundary mismatch performance of GASV, CNVer, and SVDetect, whereas DB^2, CREST, and Breakdancer are unaffected (Figure 6A). The decrease in GASV's performance can be explained by the method's conceptual use of trapezoids, determined by discordantly mapped read pairs, to define the possible boundaries of the tandem duplication. GASV finds the intersection of the trapezoids (as does DB^2) to predict the location of the tandem duplication. However, as the fragment length increases, so does the area covered by each trapezoid, causing GASV to report a larger interval for candidate start and end sites for the tandem duplication. DB^2 solves this problem by ranking the predicted start and end sites by assigning

probability values to each of them using the fragment length distribution (see Methods), and as a result does not have a deteriorating performance as the mean value of the fragment lengths increases. For similar reasons, we also observe a slight decrease in the mean boundary mismatch performance for GASV as the standard deviation of the fragment lengths increases. All other methods except SVDetect have stable mean boundary mismatch performances (Figure 6B).

Figure 6 Mean breakpoint mismatch for various levels of (A) mean value of fragment lengths, (B) standard deviation of fragment lengths, and (C) read length. Here the base call error rate, depth of coverage, duplication size are fixed at 0.01, 40X and 10 Kbp, respectively. For **(A)** and **(B)**, the read length is fixed at 75 bp. For **(A)** and **(C)**, standard deviation of the fragment lengths is fixed at 10 bp. For **(B)**, mean of the fragment lengths is 200 bp and for **(C)**, this value is fixed at 400 bp.

Lastly, we observe a poor performance for GASV in terms of mean boundary mismatch for small reads (again for similar reasons), whereas DB^2's performance is very stable for all read lengths (Figure 6C). Indeed, as the read length decreases, the area of each trapezoid induced by a discordantly aligned read pair increases. Again, we overcome this difficulty by calculating a probability value for each predicted loci pair using the empirical fragment length distribution and reporting the one with highest probability. As seen in the results of these experiments, our method is very resilient to negative effects of changing properties of the read library in terms of all metrics.

Run-time and memory consumption comparison

For each method, we computed the average time needed to produce its results, as well as its peak memory consumption on a PC that has 96 gigabytes of memory and eight Intel Xeon E5-4620 CPUs each with a clock speed of 2.20 GHz and (Table 1). Although DB^2 consumes the largest memory among all the methods, it is still tolerable when we take its superior run-time into account. It should also be taken into consideration that even today's low-end desktop computers are equipped with 8 GB of memory, which makes the memory requirement of DB^2 feasible for a high-end computer cluster used for scientific computation.

Table 1 Average run-time and memory consumption for compared methods

	DB^2	SVDetect	CNVer	Breakdancer	GASV	CREST
Run Time (seconds)	142.55	368.26	168.95	180.56	403.58	1625.092
Peak Memory Usage (kb)	8601184	5161536	4615120	144784	5309072	201024

Tandem duplications identified in two ovarian cancer genomes

To investigate whether our algorithm can identify tandem duplications in real data setting, we applied DB^2 to the paired-end read data obtained from two ovarian cancer genomes from The Cancer Genome Atlas (TCGA). The samples that we analyzed are TCGA-13-0723 and TCGA-24-0980. We identified a total of 219 tandem duplications in these genomes using our approach, which we provide in the Additional file 7: Table S1. A recent study [26] analyzing the same set of samples reported three tandem duplications – one in TCGA-13-0723 and two in TCGA-24-0980. DB^2 was able to identify these tandem duplications. In Table 2, we present the start and end sites of these duplications reported by [26] and identified by DB^2.

Table 2 Previously-reported tandem duplications identified by our method (in Hg 19 coordinates)

Sample	Chromosome	Start Bp (reported)	End Bp (reported)	Start Bp (by DB^2)	End Bp (by DB^2)
TCGA-13-0723	2	28681251	29521634	28663242	29521603
TCGA-24-0980	2	28887883	28900892	28887881	28912909
TCGA-24-0980	2	122915488	122919330	122915490	122923325

Tandem duplications identified in a melanoma genome

We also applied our method to the paired-end read data obtained from the cell line COLO-829, immortalized from a 43-year-old male with metastasis of a malignant melanoma. Illumina GAII genome analyzers were used to obtain more than 40-fold average haploid genome coverage [27]. We applied our pipeline (Figure 1) to the BAM files obtained by mapping the FASTQ-formatted paired-end read data obtained from COLO-829 cell line to the human reference genome using BWA [24]. Table 3 describes four tandem duplications (two previously reported [27] and two novel) found in this genome by DB^2. The two novel discoveries were validated with PCR (Figure 7) and Sanger sequencing (Additional file 8: Figure S7).

Table 3 Colo-829 Tandem duplications identified by our method and PCR/Sanger - validated or previously reported (Hg 18) coordinates

Chromosome	Reported*/ Sequencing Validated Start	Reported*/ Sequencing Validated End	Predicted Start Bp	Predicted End Bp	Previously Reported?*
1	222713226	222866743	222713222	222866796	Yes
7	104272303	104399536	104272363	104399571	Yes
7	114317959	114318185	114317896	114318193	No
16	80356160	80356702	80356082	80356669	No

* in the study that first sequenced this sample [27]. The two that were not previously reported are PCR (Figure 7) and Sanger Sequencing (Additional file 8: Figure S7) validated.

Figure 7 PCR results for previously unreported tandem duplications. The top panel shows the band for the PCR product generated from primers within the duplicated regions (control band), present in both COLO-829 and the NA19141 control sample (since COLO-829 is heterozygous for each duplication). In the bottom panel, the second and fourth lanes show the presence, in the COLO-829 cell line, of the third and fourth, respectively, tandem duplications given in Table 3. Lanes three and five correspond to NA19141. Here forward primers were designed left of the fusion points and reverse primers were designed right of the fusion point, creating an amplicon of about 150 bp straddling the fusion point of the duplication. See Additional file 9: Table S2 for primer sequences.

Conclusions

Tandem duplications are an important class of structural variation whose identification requires specialized algorithms. The algorithm that we propose here can identify tandem duplications with a very low false positive rate and a very low mean breakpoint mismatch (approximately 15-20 bp), even in very noisy NGS datasets, without compromising sensitivity. As shown by systematic computational experiments on simulated data, DB^2 achieves a precision of 99.6% and a recall of 77% even for an unusually noisy data (base call error rate 0.07). These results indicate that our method is not very susceptible to the effects of base calling errors in terms of making false tandem duplication predictions and false

boundary detections. One other important aspect of our algorithm is that its performance is stable even when the properties of the sequencing library or the size of tandem duplications in the target genome change. This shows the suitability of our method across NGS experiments with different characteristics.

The key to the success of DB^2 in accurate breakpoint localization is the utilization of the empirical fragment length to predict the most feasible breakpoint for a tandem duplication. As shown in Additional file 10: Figure S8, the distribution of the fragment lengths is generally not uniform in NGS experiments. Thus, given an everted (RF) read pair as the evidence for a tandem duplication, breakpoints of this duplication that indicate a higher frequency fragment length (hence higher probability for this fragment length to be observed) for this RF read pair, should have a higher probability than the others to be the real breakpoints. DB^2 uses this novel idea to precisely determine the breakpoints of the tandem duplications. Note that neither GASV, nor its extended version GASVPro employs empirical fragment length distribution to probabilistically score the potential breakpoints of structural variations. They instead assume that the lengths of all fragments are within a predefined range, and based on this assumption estimate a (rather broad) range of equally likely breakpoints for identified duplications. In contrast, we use the empirical length distribution obtained from the concordantly aligned reads to assign a probability score to each feasible breakpoint, thereby enabling ranking of candidate breakpoints in terms of their likelihood of being the correct breakpoint. As detailed in the Results and Discussion, the use of the fragment length distribution gives our method the stability for accurate boundary prediction performance.

Our method also achieves a very high precision and recall performance, substantially outperforming the SVDetect and CNVer in terms of these two measures. Although Breakdancer and GASV achieve the best precision performance among all the methods, they perform at most only 1% better than DB^2, and are substantially outperformed in terms of recall. In terms of F_1-score, our method outperforms all the other methods with increasing error rate and data coverage, showing the superiority of our method in identifying the largest set of true positive tandem duplications with the least number of false positives. Finally, the duplications identified in the two TCGA ovarian cancer samples and the COLO-829 cell line confirm the applicability of DB^2 to real datasets.

DB^2 is freely available at http://mendel.gene.cwru.edu/laframboiselab/software.php. Efforts are underway to extend the methodology to detecting non-tandem duplications, deletions and inversions.

Methods

Our method uses the BAM files that are generated by BWA [24], which aligns the FASTQ-formatted read pair files generated by the sequencer from the donor genome's (i.e. the genome under interrogation) DNA. Everted (RF) read pairs are considered to be indicative of tandem duplications [25]. The RF read pairs are those that map to the reference genome in such a way that the end with a lower mapping coordinate is aligned to the reverse strand on a chromosome, and the other end is aligned to the forward strand at a higher coordinate on the same chromosome.

Let there be M RF read pairs that map uniquely to the reference genome, and let r represent the lengths of the reads in base pairs. Note that each read pair comes from a single fragment. For each $i \in M$, let s_i and e_i denote the lowest base positions of the i^{th} pair's ends that are aligned to the reverse and forward strands, respectively (Figure 8). The standard sequencing protocol includes a size-selection step to yield fragments within a desired range with a relatively low variance. Each fragment has a length within this range, which may be considered an instance of a random variable L drawn from a distribution within this range. Thus it can be assumed that L has lower and upper bounds, denoted by $l_{\min}$ and $l_{\max}$, respectively. Let l_i denote the length of the fragment for the i^{th} RF read pair ($l_{\min} \leq l_i \leq l_{\max}$). Clearly, l_i is not observed. However, the distribution of fragment length, along with its minimum and maximum values, $l_{\min}$ and $l_{\max}$, can be determined empirically using the read pairs that are mapped to the reference genome concordantly by the alignment tool.

Figure 8 The alignment of a read pair straddling the fusion point of a tandem duplication. This figure demonstrates that the alignment of the i^{th} read pair straddling the fusion point of a tandem duplication of the region delimited by coordinates x_0 and y_0 should be everted (RF). Furthermore, the length of the i^{th} fragment should be equal to the sum of the lengths of two segments, one delimited by y_0 and e_i and the other delimited by x_0 and $s_i + r - 1$ as shown here.

Set of potential breakpoints implicated by a single discordant read pair

Suppose that there exists a tandem duplication of the segment delimited by genomic coordinates x_0 and y_0, denoted here as $t = (x_0, y_0)$. We refer to the coordinates x_0 and y_0 as respectively the start and end breakpoints of the tandem duplication t, hence (x_0, y_0) is called a breakpoint-pair. If the i^{th} fragment ($i \in M$) straddles the fusion point, then the corresponding pair is expected to have an RF discordant mapping (owing to aberrant orientation, as explained in [25]) to positions s_i and e_i on the reference genome as shown in Figure 8.

Based on the observation shown in Figure 8, the following four inequalities hold:

(i) $y_0 \geq x_0 + e_i - s_i - r - 1 + l_{\min}$,
(ii) $y_0 \leq x_0 + e_i - s_i - r - 1 + l_{\max}$,
(iii) $x_0 \leq s_i$ and
(iv) $y_0 \geq e_i + r - 1$

As seen in Figure 8, l_i is equal to the sum of the lengths of two segments in the reference genome, one delimited by y_0 and e_i and the other delimited by x_0 and $s_i + r - 1$ (i.e., $l_i = (y_0 - e_i + 1) + (s_i + r - 1 - x_0 + 1) = y_0 - x_0 - e_i + s_i + r + 1$). Since fragment length is variable, we do not know the value of l_i, but do only know its minimum and maximum possible values. Thus, we obtain $l_{\min} \leq y_0 - e_i + s_i - x_0 + r + 1 \leq l_{\max}$ which yields to the inequalities (i) and (ii). Furthermore, the two reads will flank the fusion point but not contain it. These two restrictions are expressed by the inequalities (iii) and (iv).

Therefore, given the mapping of the i^{th} RF read pair (i.e., e_i and s_i) and the minimum and maximum values of the fragment length, $l_{\min}$ and $l_{\max}$, we can define the range of possible start and end breakpoints of the tandem duplication that induce the i^{th} discordant mapping using the inequalities (i), (ii), (iii) and (iv). The inequalities geometrically define a trapezoid in CxC plane, where C represents the coordinates of the reference chromosome. This idea was introduced by [14] for the identification of various types of structural variations. The

trapezoid (shown in Figure 9 as the light blue region) comprises the set of all possible pairs of start and end breakpoints (x, y) delimiting a tandem duplication that can potentially induce the i^{th} RF read pair. We denote the set of breakpoint-pairs in this trapezoid as W. More formally,

$$W = \{(x, y) \in CxC : (y \leq x + e_i - s_i - r - 1 + l_{max}) \wedge (y \geq x + e_i - s_i - r - 1 + l_{min}) \wedge (x \leq s_i) \wedge (y \geq e_i + r - 1)\}.$$

Figure 9 The geometric representation of the set of all potential pairs of start and end breakpoint coordinates. In this figure, the light blue region denoted by W represents the set of all potential pairs of start and end breakpoint coordinates of a tandem duplication inducing an RF read pair that aligns to (s_i, e_i).

Detecting distinct putative tandem duplications

A donor genome will often harbor multiple tandem duplications. Furthermore, as depth coverage for a typical experiment increases, one would expect that more than one read pair straddling the fusion point of each tandem duplication will be produced during the sequencing of a donor genome. This gives us the opportunity to use multiple read pairs to predict the breakpoints of the tandem duplications more precisely because we have more statistical power and more information as more RF read pairs are induced by the same tandem duplication. However, this also necessitates the identification of multiple read pairs that are induced by the same tandem duplication.

Given M, r, $l_{\min}$ and $l_{\max}$, we can take advantage of the fact that, if two RF read pairs i and j are induced by the same tandem duplication (for ease of notation, we now denote each read pair by its corresponding index), then the real coordinates of that duplication should lie in the intersection of the corresponding trapezoids W_i and W_j. It follows that a tandem duplication in the donor genome can be identified by finding the maximum subset, denoted by S, of the set of all aligned RF read pairs such that $\cap_{i \in S} W_i \neq \emptyset$ (i.e. all trapezoids corresponding to read pairs in S intersect in at least one point). In this case, we say that the tandem duplication t induces the RF pair set S. Thus, the problem of discovering multiple tandem duplications can be framed as the problem of finding the set $\mathbf{S} = \{S_1, S_2, \ldots, S_n\}$ where each read pair set $S_k \in \mathbf{S}$ is induced by a unique tandem duplication t_k.

In an ideal setting, two trapezoids associated with distinct sets S_q and S_p ($q \neq p$) should not overlap, since no read pair can straddle two tandem duplications simultaneously (assuming that the tandem duplications do not overlap). Thus $\mathbf{S}$ is ideally a partitioning of the set of all RF read pairs into disjoint subsets (i.e., $\cup_{S_k \in \mathbf{S}} S_k = M$ and $S_q \cap S_k = \emptyset$ for all $q \neq k$) such that all read pairs in each S_k have corresponding trapezoids intersecting at least one point, and trapezoids corresponding to read pairs from two different S's do not intersect. However, noisy sequence data (e.g. base call or alignment errors) can lead to imperfect partitioning of the read pair set. As such, we relax the condition requiring that the trapezoids induced by the same tandem duplication contain the breakpoint coordinates of duplication. Instead, we require that there is a mutual intersection between the trapezoids induced by the same duplication. Formally, we require that each S_k satisfies the condition: $\forall i \in S_k$, $\exists j \in S_k$, such that $i \neq j$ and $W_i \cap W_j \neq \emptyset$.

An important step in our method for finding the partitioning **S** involves determining which trapezoids intersect a given trapezoid. To perform this operation quickly, we implement an R* tree [28] data structure, which is a variant of the R tree data structure [29] used for indexing spatial information. R-trees are hierarchical data structures, which are used for the dynamic organization of a set of multi-dimensional geometric objects by representing them with the minimum bounding multi-dimensional rectangles. DB^2 builds an R* tree using the Java implementation freely available at [30] to index all of the trapezoids of *M*, and uses this data tree to identify the trapezoids that intersect a given trapezoid. In our experimental evaluation, we have observed that using R* trees for intersection identification is computationally more efficient compared to a naive method, which would check all the trapezoids in *M* for intersection.

To find the disjoint sets of intersecting trapezoids, we use a method similar to that used for finding the connected components of an undirected graph [31]. Namely, we implement a breadth-first search (BFS) like algorithm, which starts with an arbitrary trapezoid, *i*, finds all trapezoids that intersect with *i*, and then iteratively finds all trapezoids that intersect with these trapezoids. This procedure discovers the entire connected trapezoid set containing *i* before it returns. Next, it assigns the newly found connected trapezoid set into a set S_k (where initially k =1) and *M* is updated as $M = M \setminus S_k$ and $k = k + 1$. Then the same procedure is repeated for the updated *M* until *M* becomes empty. The set of tandem duplications, $T = \{t_1, t_2, \ldots, t_n\}$ corresponding to the set $\mathbf{S} = \{S_1, S_2, \ldots, S_n\}$ of connected trapezoids represents our algorithm's final set of predicted tandem duplications. At this stage, the tandem duplication breakpoints are not yet precisely defined. Optimally determining these breakpoints is the next step.

Set of potential breakpoints implicated by multiple discordant read pairs

After we determine the set of distinct tandem duplications, T, and the set, S_k, of RF read pairs induced by each tandem duplication, the next step is to estimate the start and end breakpoint sites of each t_k. Ideally, the set of candidate breakpoints would be the intersection of all trapezoids corresponding to the read pairs in S_k . However, due to sequencing and mapping errors, this intersection is often empty. For this reason, we consider the set of breakpoints that are supported by the maximum number of RF pairs as candidate breakpoints. In other words, we define Ω_k as the set of all coordinates in the CxC plane that are contained by the maximum number of trapezoids corresponding to read pairs in S_k. The set Ω_k for each t_k is the set of candidate breakpoint-pair coordinates for the corresponding tandem duplication.

Scoring candidate breakpoints based on the observed distribution of fragment length

Once we identify the set of candidate breakpoint-pairs for each tandem duplication, the final step is to score and rank these candidate breakpoint-pairs. For this purpose, we introduce a probabilistic model that makes use of the empirical distribution of fragment length.

In order to motivate the proposed approach, we first consider the case when only a single RF read pair, say the i^{th} pair, is induced by a tandem duplication. Recall that W_i denotes the set of all possible genomic coordinates delimiting the tandem duplication that induces the i^{th} RF read pair. Now define $P[(x, y) \mid i]$ (where $(x, y) \in W_i$) as the probability of this tandemly duplicated segment being delimited by base positions x and y, given only the i^{th} RF read pair and the empirical fragment length distribution. If the distribution of fragment length, *L*, was

uniform, then all the genomic coordinates in W_i would have the same probability of being the true breakpoint-pairs. However, in practice, we know that fragment length is not uniformly distributed. This can be seen, for example, in the COLO-829 cell line data [27] (Additional file 10: Figure S8).

Each candidate breakpoint-pair $(x, y) \in W_i$ corresponds to a specific fragment length, since for breakpoint-pair (x, y), the corresponding fragment length can be computed as $y - e_i + s_i - x + r + 1$. Therefore, applying Bayes' theorem, we can conclude that the probability score for each coordinate pair in W_i is proportional to the probability that the i^{th} fragment has the corresponding length. Consequently, we can compute the probability of the i^{th} RF read pair being induced by a tandem duplication of the genomic segment delimited by coordinates x and y as:

$$P\left[(x,y) \mid i\right] = \frac{\sigma_i(x,y)}{\sum_{(a,b)\in W_i}\sigma_i(a,b)}$$

where $\sigma_i(x, y) = P_L\ [L = y - e_i + s_i - x + r + 1]$ is based on the empirical fragment length distribution.

Now we generalize this observation to the case where a tandem duplication is supported by multiple RF read pairs. For each $(x, y) \in \Omega_k$, let $Z_{(x, y)}$ denote the set of RF read pairs that support the candidate breakpoint-pair (x, y), i.e., the trapezoids for these RF read pairs contain (x, y). Assuming that the lengths of different fragments are independent, the probability of $(x, y) \in \Omega_k$ being the start and end breakpoint-pair of the k^{th} tandem duplication will be proportional to the product of the probabilities of observing the corresponding fragment lengths of the read pairs in $Z_{(x, y)}$. Thus, we can compute the probability, denoted by $P[(x, y) \mid S_k]$, that a point $(x, y) \in \Omega_k$ is the real breakpoint-pair of t_k as follows:

$$P\left[(x,y) \mid S_k\right] = \begin{cases} \dfrac{\prod_{j \in Z_{(x,y)}} \sigma_j(x,y)}{\sum_{(a,b)\in\Omega_k}\prod_{j\in Z_{(a,b)}}\sigma_j(a,b)} & (x,y)\in\Omega_k \\ 0 & \textit{otherwise} \end{cases}$$

After computing this probability score for each $(x, y) \in \Omega_k$, we report the (x, y) with the highest probability as the predicted breakpoint-pair of t_k (in the case of a tie, the point is randomly selected from those with highest probability). Formally t_k is defined as:

$$t_k = argmax_{(x,y)\in\Omega_k}\left(P\left[(x,y) \mid S_k\right]\right).$$

As an example, Figure 10 shows the probability distribution computed by our algorithm for a simulated tandem duplication on human reference chromosome 22, which induces three RF read pairs shown with three trapezoids. In this case, S consists of only these three RF read pairs. Notice that the real breakpoint coordinate of this tandem duplication, shown by "X", lies in the common intersection of these three trapezoids, Ω.

Figure 10 A heatmap representation of the probability scores of the potential breakpoint coordinates of an example tandem duplication. In this figure, we show a heatmap of the probability scores of the potential breakpoint coordinates of an example tandem duplication with true start and end breakpoints, (31219230, 31224279) on chromosome 22. In this case, *S* contains three read pairs shown by the dotted trapezoids and Ω contains only the points in the core area for which a probability score is computed.

Conflict resolution among tandem duplications

After the set of all distinct tandem duplications, T, is identified along with their coordinates, it is possible that some of the predicted duplications overlap with each other in terms of their boundaries. In such a case, we say that the tandem duplications are conflicting with each other and the conflict is likely caused by false positive tandem duplications that are the results of the noisy data. Therefore, a conflict resolution procedure is needed to find the subset of the tandem duplications out of T, containing only non-overlapping duplications that are possibly the true positives. Toward this end, we employ a simple idea based on the maximum parsimony principle. Namely, we assume that the true tandem duplications existing in a donor genome do not overlap; hence, the duplications that overlap with most of the other predicted duplications are falsely identified.

To obtain the true positive set, we use a greedy approach. Starting with T, we eliminate the tandem duplication that overlaps with most of the duplications in T to obtain a subset T′ of T. We then check if there is still any conflict in the new set of tandem duplications, T′. If there is no conflict, DB^2 reports T′ as the final set of tandem duplications. Otherwise, the procedure is iterated until there is no conflict left.

Data generation for simulation experiments

We have implemented a freely available NGS data generator [32]. Our data generator first selects a user-defined number of base positions uniformly at random on the reference chromosome provided by the user. These randomly selected positions mark the starting point of each tandem duplication. Next, the size of each duplication is drawn from a normal distribution, whose mean and standard deviation are defined by the user. For our simulations, we have used 10 Kbp and 100 bp as the default mean and standard deviation, respectively, and simulate 1000 tandem duplications for each experiment. After determining the start and end breakpoint-pair for each duplication, our data generator inserts an exact copy of the genomic segment delimited by these two coordinates, right after the end breakpoint to spike in the tandem duplication.

We then select a user-defined number (which is computed according to the user-defined depth of coverage) of base positions $v_1, v_2, \ldots, v_u$ on the genome as the start location of each read pair. Subsequently, left and right ends of the i^{th} read pair are generated as follows. A "read" of r bases (in the current study, we use $r = 75$ as the default value of read length) starting from selected base position is extracted from the reference genome in the forward direction. This sequence forms the left end of the read pair. For generating the right end, our simulator first selects an l_i value from a normal distribution L (with a default mean value of 200 and default standard deviation of 10). Note that the empirical length distribution of the paired-end reads obtained from the COLO-829 cell line [27] is similar to this setting. The start locus of the right end on the reverse strand is determined as $v_i + l_i$. The right end read is formed by reading r bases of the reverse strand of the genome in the reverse direction (i.e.,

read direction is from right to left and the bases in the right end sequence are the complementary bases of the forward strand of the genome). During the read generation process, we replace the base at each locus with a randomly selected base with a user-defined probability value (i.e., base call error rate) to simulate the sequencing errors.

Additional data files

The following additional data are available with the online version of this paper. Additional data file 1 is a figure showing the mean number of supporting read pairs at various levels of base calling error rate. Additional data file 2 is a figure demonstrating the Precision, Recall and F1-score performances of the methods at different duplications sizes. Additional data file 3 is a figure showing the mean breakpoint mismatch performances at different duplication sizes. Additional data files 4, 5 and 6 are figures demonstrating the Precision, Recall and F_1-score performances of the methods at different levels of fragment length mean/standard deviation and read length, respectively. Additional data file 7 is a table listing the genomic location of the tandem duplications identified in TCGA ovarian cancer samples. Additional data file 8 is a table listing the PCR primer sequences used for validating the two novel tandem duplications discovered by DB^2 in COLO-829 cell line. Additional data file 9 contains two figures demonstrating the Sanger validation of previously unreported tandem duplications in COLO-829. Additional data file 10 is a figure that shows the empirical fragment length distribution of sequenced COLO-829 cell line.

Abbreviations

NGS, Next-generation sequencing; BAM, Binary Alignment/Map; PCR, Polymerase chain reaction; BFS, Breadth-first search

Competing interests

The authors declare that they have no competing interests

Authors' contributions

GY, MK and TL designed the algorithms. GY implemented the DB^2 framework and collected the results for analysis and analyzed the results. MPG and SM performed the PCR and sequencing validation. All authors read and approved the final manuscript.

Acknowledgements

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Awards R25-CA094186, R01-CA131341, the National Science Foundation under Award IIS-0916102, and the American Cancer Society under award 123436-RSG-12-159-01-DMC.

References

1. McBride DJ, Etemadmoghadam D, Cooke SL, Alsop K, George J, Butler A, Cho J, Galappaththige D, Greenman C, Howarth KD, Lau KW, Ng CK, Raine K, Teague J, Wedge DC, Cancer Study Group AO, Caubit X, Stratton MR, Brenton JD, Campbell PJ, Futreal PA, Bowtell DD: **Tandem duplication of chromosomal segments is common in ovarian and breast cancer genomes.** *J Pathol* 2012, **227:**446–455.

2. Nakao M, Yokota S, Iwai T, Kaneko H, Horiike S, Kashima K, Sonoda Y, Fujimoto T, Misawa S: **Internal tandem duplication of the flt3 gene found in acute myeloid leukemia.** *Leukemia* 1996, **10:**1911–1918.

3. Yokota S, Kiyoi H, Nakao M, Iwai T, Misawa S, Okuda T, Sonoda Y, Abe T, Kahsima K, Matsuo Y, Naoe T: **Internal tandem duplication of the FLT3 gene is preferentially seen in acute myeloid leukemia and myelodysplastic syndrome among various hematological malignancies. A study on a large series of patients and cell lines.** *Leukemia* 1997, **11:**1605–1609.

4. Schichman SA, Caligiuri MA, Gu Y, Strout MP, Canaani E, Bloomfield CD, Croce CM: **ALL-1 partial duplication in acute leukemia.** *Proc Natl Acad Sci U S A* 1994, **91:**6236–6239.

5. Yuksel-Apak M, Bögershausen N, Pawlik B, Li Y, Apak S, Uyguner O, Milz E, Nürnberg G, Karaman B, Gülgören A, Grzeschik KH, Nürnberg P, Kayserili H, Wollnik B: **A large duplication involving the IHH locus mimics acrocallosal syndrome.** *Eur J Hum Genet* 2012, **20:**639–644.

6. Naik S, Thomas NS, Davies JH, Lever M, Raponi M, Baralle D, Temple IK, Caliebe A: **Novel tandem duplication in exon 1 of the SNURF/SNRPN gene in a child with transient excessive eating behaviour and weight gain.** *Mol Syndromol* 2012, **2:**76–80.

7. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M: **Paired-end mapping reveals extensive structural variation in the human genome.** *Science* 2007, **318:**420–426.

8. Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, Stebbings LA, Leroy C, Edkins S, Hardy C, Teague JW, Menzies A, Goodhead I, Turner DJ, Clee CM, Quail MA, Cox A, Brown C, Durbin R, Hurles ME, Edwards PA, Bignell GR, Stratton MR, Futreal PA: **Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing.** *Nat Genet* 2008, **40:**722–729.

9. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, *et al*: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature* 2008, **456:**53–59.

10. Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z, Snyder M, Gerstein MB: **PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data.** *Genome Biol* 2009, **10:**R23.

11. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, Ding L, Mardis ER: **BreakDancer: an algorithm for high-resolution mapping of genomic structural variation.** *Nat Methods* 2009, **6:**677–681.

12. Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC: **Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes.** *Genome Res* 2009, **19:**1270–1278.

13. Hormozdiari F, Hajirasouliha I, Dao P, Hach F, Yorukoglu D, Alkan C, Eichler EE, Sahinalp SC: **Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery.** *Bioinformatics* 2010, **26:**350–357.

14. Sindi S, Helman E, Bashir A, Raphael BJ: **A geometric approach for classification and comparison of structural variants.** *Bioinformatics* 2009, **25:**222–230.

15. Sindi SS, Onal S, Peng LC, Wu HT, Raphael BJ: **An integrative probabilistic model for identification of structural variation in sequencing data.** *Genome Biol* 2012, **13:**R22.

16. Wang J, Mullighan CG, Easton J, Roberts S, Heatley SL, Ma J, Rusch MC, Chen K, Harris CC, Ding L, Holmfeldt L, Payne-Turner D, Fan X, Wei L, Zhao D, Obenauer JC, Naeve C, Mardis ER, Wilson RK, Downing JR, Zhang J: **CREST maps somatic structural variation in cancer genomes with base-pair resolution.** *Nat Methods* 2011, **8:**652–654.

17. Hajirasouliha I, Hormozdiari F, Alkan C, Kidd JM, Birol I, Eichler EE, Sahinalp SC: **Detection and characterization of novel sequence insertions using paired-end next-generation sequencing.** *Bioinformatics* 2010, **26:**1277–1283.

18. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J: **Sensitive and accurate detection of copy number variants using read depth of coverage.** *Genome Res* 2009, **19:**1586–1592.

19. Chiang DY, Getz G, Jaffe DB, O'Kelly MJ, Zhao X, Carter SL, Russ C, Nusbaum C, Meyerson M, Lander ES: **High-resolution mapping of copy-number alterations with massively parallel sequencing.** *Nat Methods* 2009, **6:**99–103.

20. Zeitouni B, Boeva V, Janoueix-Lerosey I, Loeillet S, Legoix-né P, Nicolas A, Delattre O, Barillot E: **SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data.** *Bioinformatics* 2010, **26:**1895–1896.

21. Medvedev P, Fiume M, Dzamba M, Smith T, Brudno M: **Detecting copy number variation with mated short reads.** *Genome Res* 2010, **20:**1613–1622.

22. 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467:**1061–1073.

23. Qi J, Zhao F: **inGAP-sv: a novel scheme to identify and visualize structural variation from paired end mapping data.** *Nucleic Acids Res* 2011, **39**(Web Server issue)**:**W567–W575.

24. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2010, **26:**589–595.

25. Alkan C, Coe BP, Eichler EE: **Genome structural variation discovery and genotyping.** *Nat Rev Genet* 2011, **12:**363–376.

26. Oesper L, Ritz A, Aerni SJ, Drebin R, Raphael BJ: **Reconstructing cancer genomes from paired-end sequencing data.** *BMC Bioinforma* 2012, **13**(6)**:**S10.

27. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin ML, Ordóñez GR, Bignell GR, Ye K, Alipaz J, Bauer MJ, Beare D, Butler A, Carter RJ, Chen L, Cox AJ, Edkins S, Kokko-Gonzales PI, Gormley NA, Grocock RJ, Haudenschild CD, Hims MM, James T, Jia M, Kingsbury Z, Leroy C, Marshall J, Menzies A, *et al*: **A comprehensive catalogue of somatic mutations from a human cancer genome.** *Nature* 2010, **463:**191–196.

28. Beckmann N, Kriegel HP, Schneider R, Seeger B: **The R*-tree: an efficient and robust access method for points and rectangles**. In *Proceedings of the ACM SIGMOD: May 23-25, 1990.* Edited by Hector G-M, Jagadish HV. Atlantic City: ACM Press; 1990:322–331.

29. Guttman A: **R-Trees: a dynamic index structure for spatial searching**. In *Proceedings of the ACM SIGMOD.* Edited by Beatrice Yormark. Boston: ACM Press; 1984:47–57.

30. *R* tree source code download page.* http://www.rtreeportal.org/code/Rstar-java.zip.

31. Hopcroft J, Tarjan R: **Efficient algorithms for graph manipulation.** *Commun ACM* 1973, **16:**372–378.

32. *LaFramboise Laboratory Software Website.* http://mendel.gene.cwru.edu/laframboiselab/software.php.

Additional files

Additional_file_1 as PDF

Additional file 1: Figure S1 Mean number of supporting reads at various levels of base calling error.

Additional_file_2 as PDF

Additional file 2: Figure S2 Performance as a function of duplication size.

Additional_file_3 as PDF

Additional file 3: Figure S3 Mean Breakpoint Mismatch as a function of duplication sizes.

Additional_file_4 as PDF

Additional file 4: Figure S4 Performance as a function of fragment length.

Additional_file_5 as PDF

Additional file 5: Figure S5 Performance as a function of standard deviation of fragment lengths.

Additional_file_6 as PDF

Additional file 6: Figure S6 Performance as a function of read length.

Additional_file_7 as XLS

Additional file 7: Table S1 List of identified tandem duplications.

Additional_file_8 as PDF

Additional file 8: Figure S7 Sanger validation of novel tandem duplications.

Additional_file_9 as XLSX

Additional file 9: Table S2 Primer Sequences.

Additional_file_10 as PDF

Additional file 10: Figure S8 Empirical fragment length distribution of COLO-829.

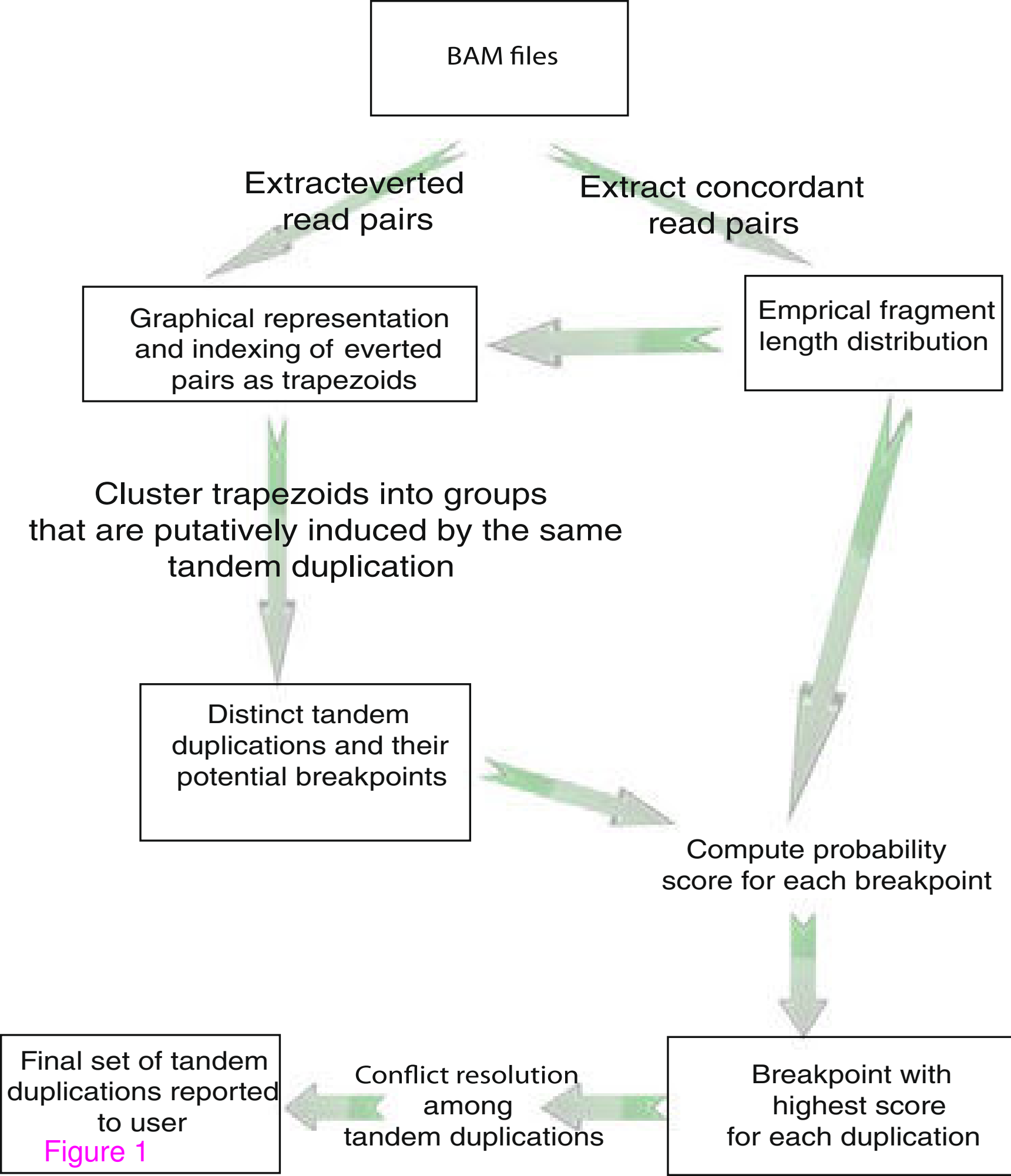
BAM files
Extracteverted read pairs
Extract concordant read pairs
Graphical representation and indexing of everted pairs as trapezoids
Emprical fragment length distribution
Cluster trapezoids into groups that are putatively induced by the same tandem duplication
Distinct tandem duplications and their potential breakpoints
Compute probability score for each breakpoint
Final set of tandem duplications reported to user
Conflict resolution among tandem duplications
Breakpoint with highest score for each duplication

Figure 1

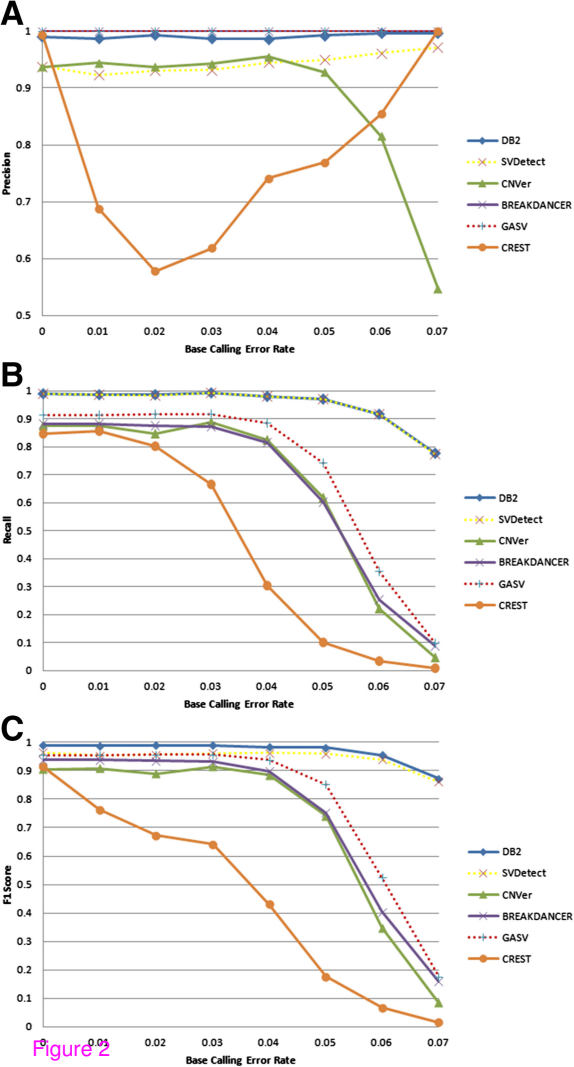
A
Precision
1
0.9
0.8
0.7
0.6
0.5
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Base Calling Error Rate
DB2
SVDetect
CNVer
BREAKDANCER
GASV
CREST
B
Recall
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Base Calling Error Rate
DB2
SVDetect
CNVer
BREAKDANCER
GASV
CREST
C
F1Score
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Base Calling Error Rate
DB2
SVDetect
CNVer
BREAKDANCER
GASV
CREST

Figure 2

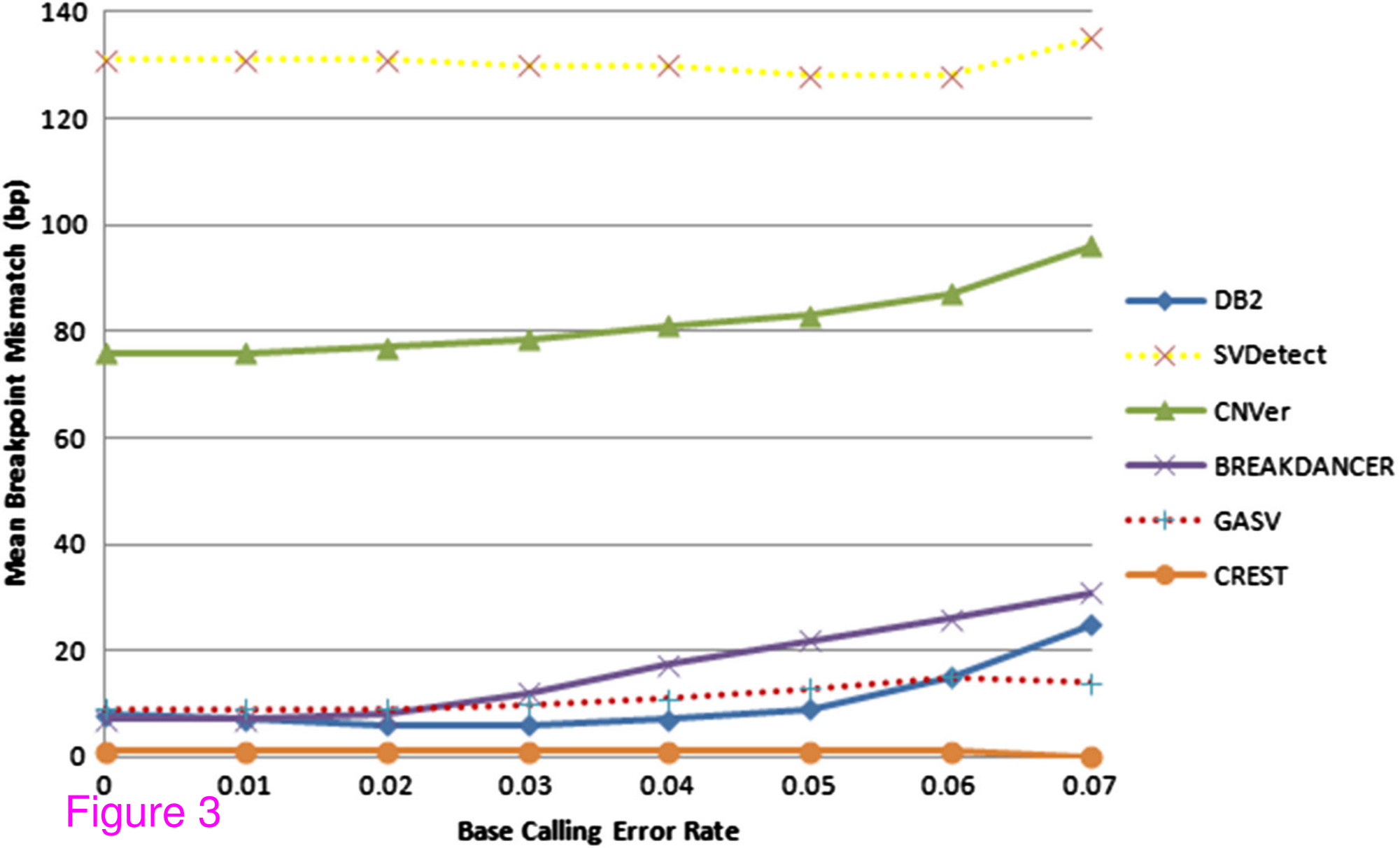
Mean Breakpoint Mismatch (bp)
140
120
100
80
60
40
20
0
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Base Calling Error Rate
DB2
SVDetect
CNVer
BREAKDANCER
GASV
CREST

Figure 3

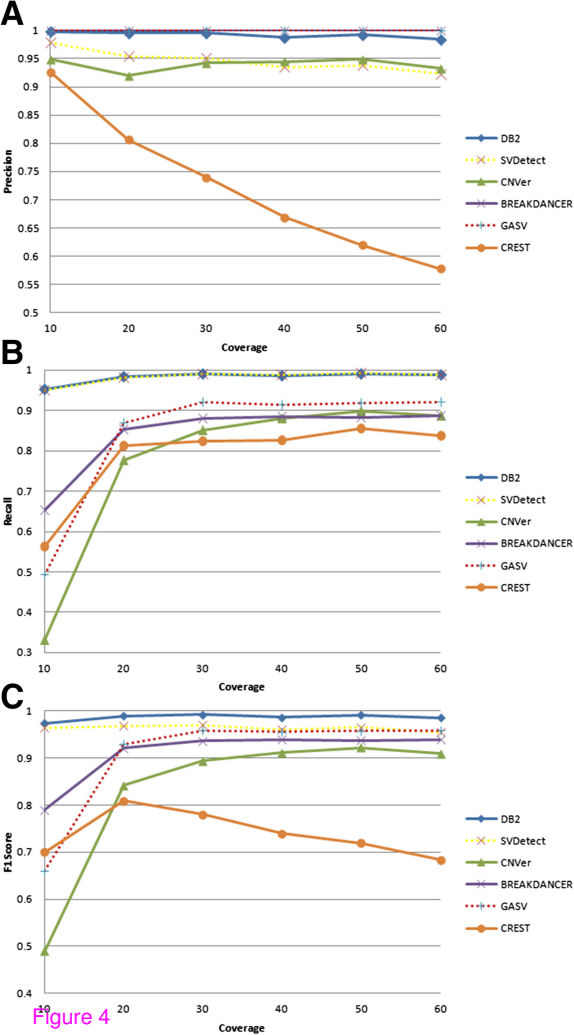
A
Precision
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
10
20
30
40
50
60
Coverage
DB2
SVDetect
CNVer
BREAKDANCER
GASV
CREST
B
Recall
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
10
20
30
40
50
60
Coverage
DB2
SVDetect
CNVer
BREAKDANCER
GASV
CREST
C
F1score
1
0.9
0.8
0.7
0.6
0.5
0.4
10
20
30
40
50
60
Coverage
DB2
SVDetect
CNVer
BREAKDANCER
GASV
CREST

Figure 4

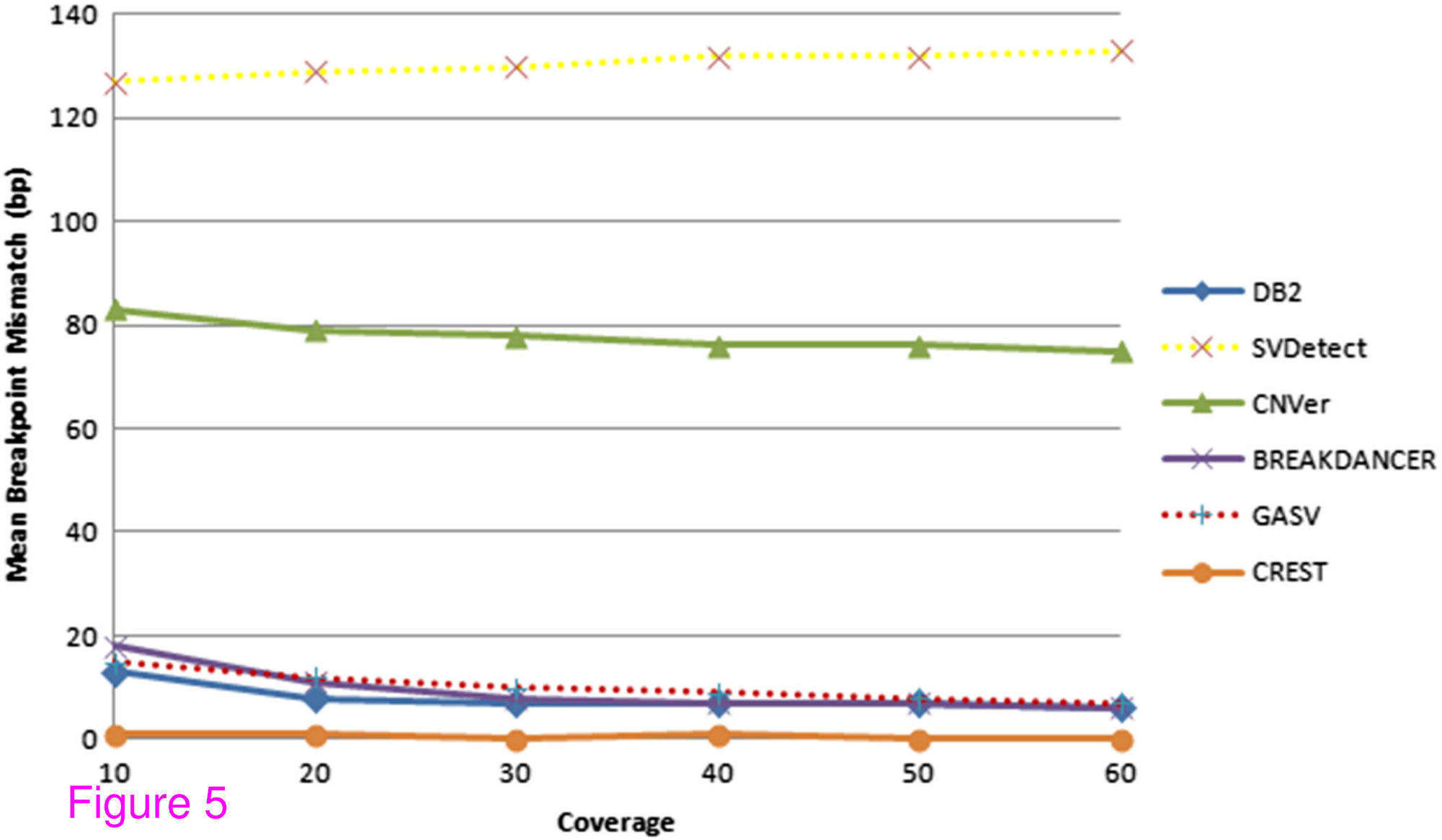
Mean Breakpoint Mismatch (bp)
140
120
100
80
60
40
20
0
10
20
30
40
50
60
Coverage
DB2
SVDetect
CNVer
BREAKDANCER
GASV
CREST

Figure 5

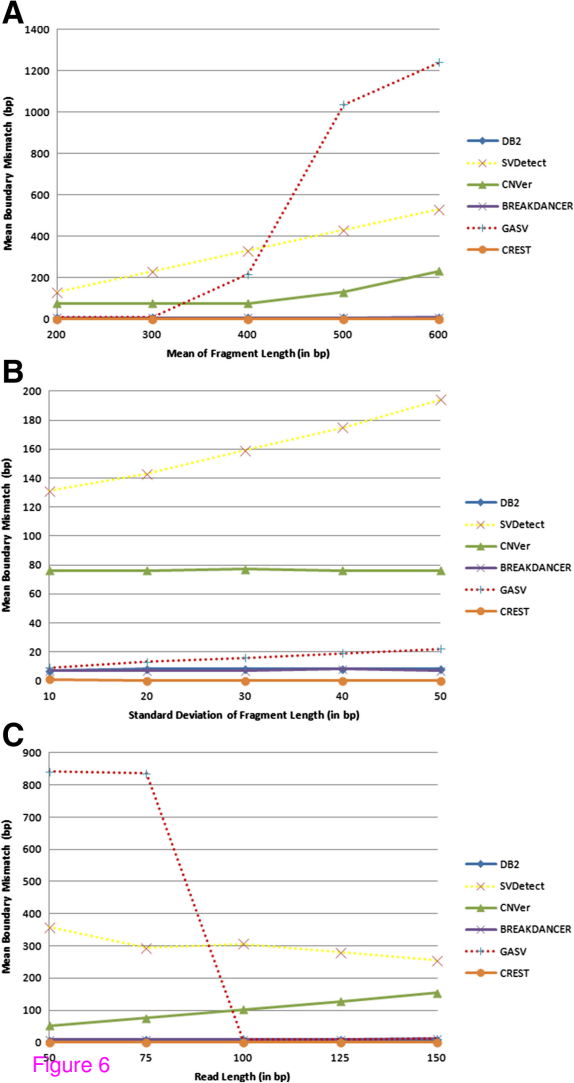
A
Mean Boundary Mismatch (bp)
1400
1200
1000
800
600
400
200
0
200
300
400
500
600
Mean of Fragment Length (in bp)
DB2
SVDetect
CNVer
BREAKDANCER
GASV
CREST
B
Mean Boundary Mismatch (bp)
200
180
160
140
120
100
80
60
40
20
0
10
20
30
40
50
Standard Deviation of Fragment Length (in bp)
DB2
SVDetect
CNVer
BREAKDANCER
GASV
CREST
C
Mean Boundary Mismatch (bp)
900
800
700
600
500
400
300
200
100
0
50
75
100
125
150
Read Length (in bp)
DB2
SVDetect
CNVer
BREAKDANCER
GASV
CREST

Figure 6

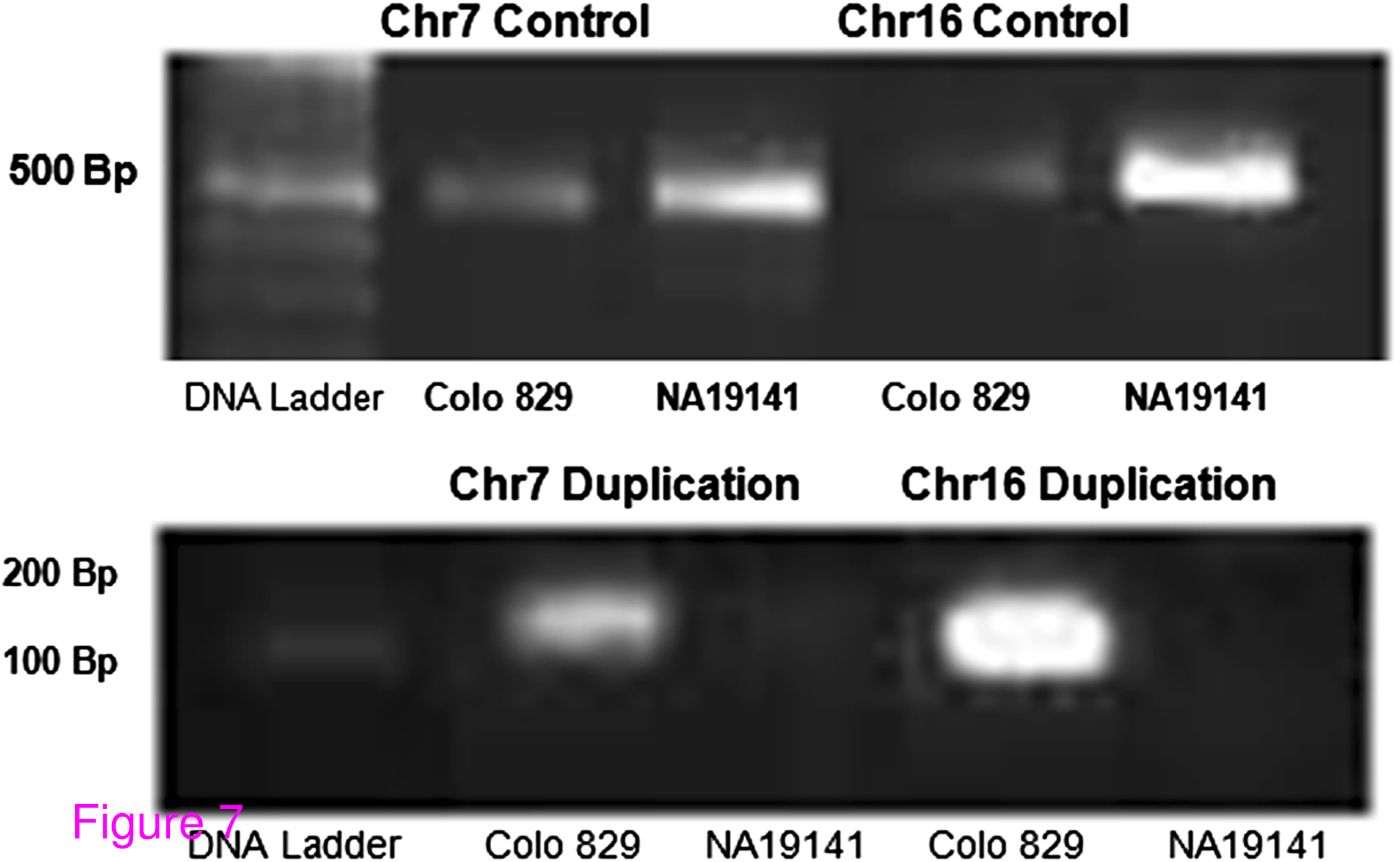
Chr7 Control
Chr16 Control
500 Bp
DNA Ladder
Colo 829
NA19141
Colo 829
NA19141
Chr7 Duplication
Chr16 Duplication
200 Bp
100 Bp
DNA Ladder
Colo 829
NA19141
Colo 829
NA19141

Figure 7

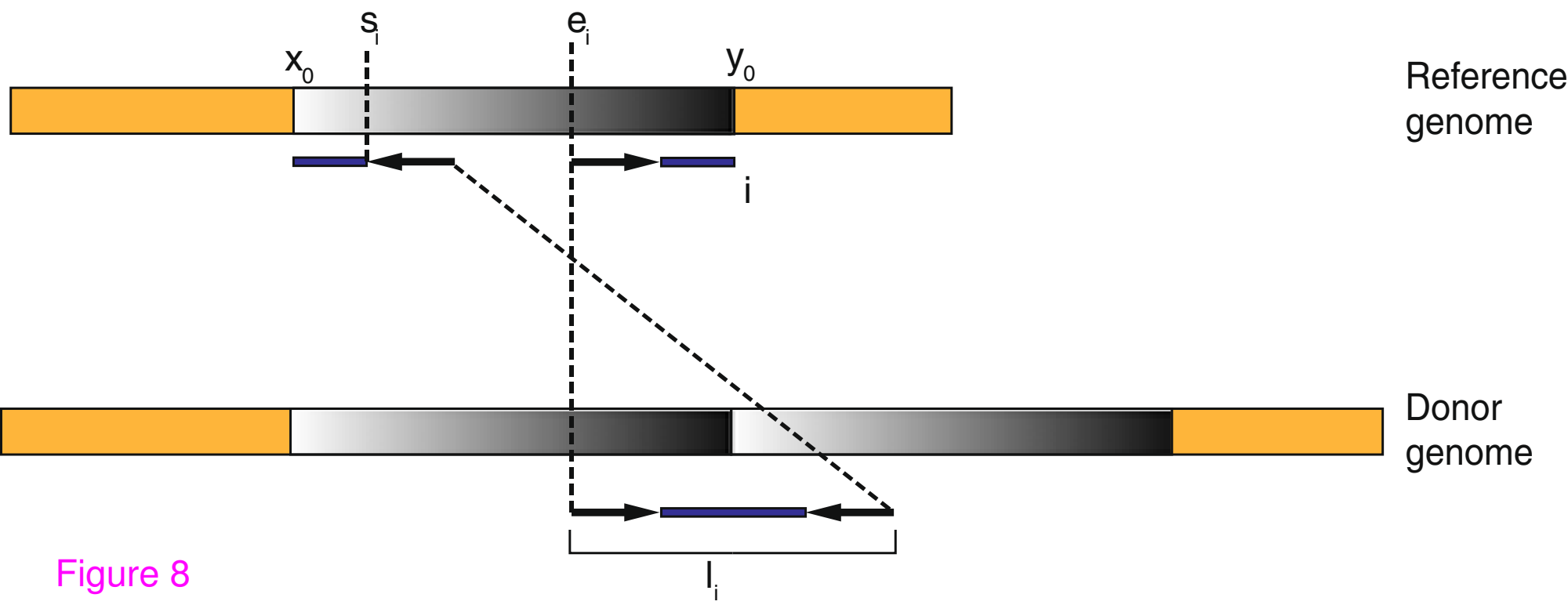

Figure 8

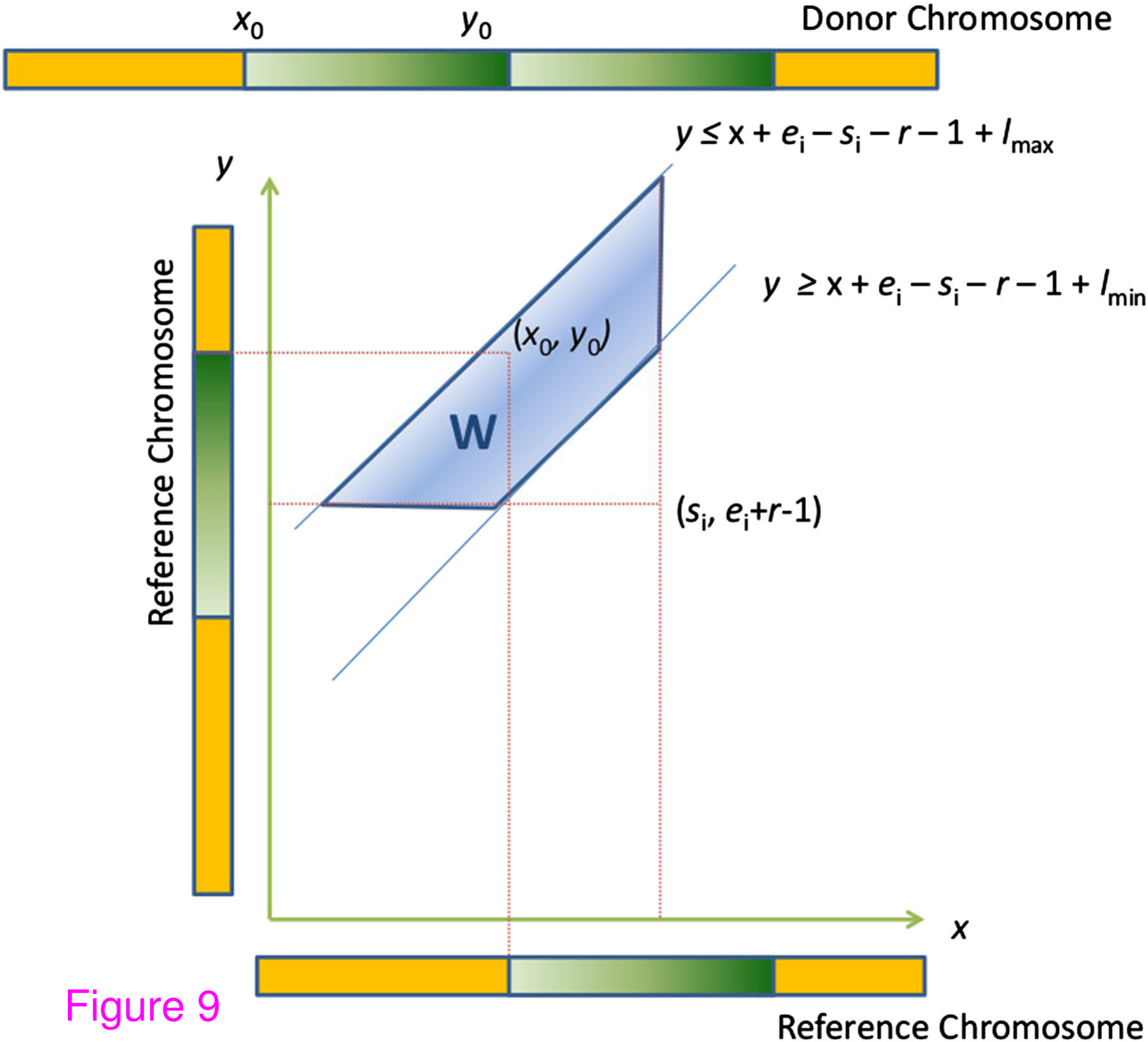
x_0
y_0
Donor Chromosome
$y \le x + e_i - s_i - r - 1 + l_{max}$
$y \ge x + e_i - s_i - r - 1 + l_{min}$
y
Reference Chromosome
(x_0, y_0)
W
(s_i, e_i+r-1)
x
Reference Chromosome

Figure 9


Chromosome 22 base position
31219350
31224300
31224250
31219200
31219250
31219300
Chromosome 22 base position

Figure 10

Additional files provided with this submission:

Additional file 1: 1333461883101581_add1.pdf, 217K
http://www.biomedcentral.com/imedia/1815704182123042/supp1.pdf
Additional file 2: 1333461883101581_add2.pdf, 823K
http://www.biomedcentral.com/imedia/1323358340123042/supp2.pdf
Additional file 3: 1333461883101581_add3.pdf, 250K
http://www.biomedcentral.com/imedia/1163931430123042/supp3.pdf
Additional file 4: 1333461883101581_add4.pdf, 821K
http://www.biomedcentral.com/imedia/1298515305123042/supp4.pdf
Additional file 5: 1333461883101581_add5.pdf, 848K
http://www.biomedcentral.com/imedia/2979968891230429/supp5.pdf
Additional file 6: 1333461883101581_add6.pdf, 819K
http://www.biomedcentral.com/imedia/2140207365123042/supp6.pdf
Additional file 7: 1333461883101581_add7.xls, 28K
http://www.biomedcentral.com/imedia/5009683151230429/supp7.xls
Additional file 8: 1333461883101581_add8.pdf, 71K
http://www.biomedcentral.com/imedia/1822266593123042/supp8.pdf
Additional file 9: 1333461883101581_add9.xlsx, 9K
http://www.biomedcentral.com/imedia/8798100731230429/supp9.xlsx
Additional file 10: 1333461883101581_add10.pdf, 5566K
http://www.biomedcentral.com/imedia/1090101177123042/supp10.pdf