Published online 1 February 2016

Nucleic Acids Research, 2016, Vol. 44, No. 8 ***e78***
doi: 10.1093/nar/gkw026

COSMOS: accurate detection of somatic structural variations through asymmetric comparison between tumor and normal samples

Koichi Yamagata[1], Ayako Yamanishi[2], Chikara Kokubu[2], Junji Takeda[2] and Jun Sese[1,3,*]

[1]Biotechnology Research Institute for Drug Discovery, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, 135-0064, Japan, [2]Department of Genome Biology, Graduate School of Medicine, Osaka University, Osaka, 565-0871, Japan and [3]Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, 135-0064, Japan

Received September 13, 2015; Revised January 01, 2016; Accepted January 11, 2016

ABSTRACT

An important challenge in cancer genomics is precise detection of structural variations (SVs) by high-throughput short-read sequencing, which is hampered by the high false discovery rates of existing analysis tools. Here, we propose an accurate SV detection method named COSMOS, which compares the statistics of the mapped read pairs in tumor samples with isogenic normal control samples in a distinct asymmetric manner. COSMOS also prioritizes the candidate SVs using strand-specific read-depth information. Performance tests on modeled tumor genomes revealed that COSMOS outperformed existing methods in terms of F-measure. We also applied COSMOS to an experimental mouse cell-based model, in which SVs were induced by genome engineering and gamma-ray irradiation, followed by polymerase chain reaction-based confirmation. The precision of COSMOS was 84.5%, while the next best existing method was 70.4%. Moreover, the sensitivity of COSMOS was the highest, indicating that COSMOS has great potential for cancer genome analysis.

INTRODUCTION

Genomic structural variations (SVs), such as deletions, inversions, translocations and duplications, are a major source of genetic diversity in both cancers (1,2) and inherited diseases (3–5). Many researchers have tried to uncover the association between SVs and such disorders (6,7). Recent studies using high-throughput sequencing revealed that the frequency and complexity of SVs occurring in somatic cancerous cells are much higher than previously expected (8–12). Therefore, the development of a highly sensitive and accurate SV detection method has been widely anticipated.

The accurate detection of SVs in tumor cells is both computationally and statistically difficult to achieve (13,14). To find somatic SVs, SV detection methods (15–23) are usually applied to tumor and normal samples independently, followed by subsequent comparison of the results. However, this procedure often generates many false discoveries from sequencing errors and polymorphic differences between the samples and reference genomes. Furthermore, tumor tissues are often heterogeneous (24,25) and only a small percentage of the cells in a tumor have SVs, making the data analysis more difficult. The high false-positive rate of SV detection methods has prevented efficient processing and better understanding of high-throughput sequencing data to elucidate the association between SVs and tumorigenesis.

Direct comparison of tumor and normal samples might reduce the false discovery rate. LUMPY (18) can detect SVs from multiple samples simultaneously and easily compares the SVs between the samples. However, its assumption that two or more SVs do not overlap might cause a problem if they are used to analyze complex SVs such as chromothripsis (8,9,26). Somatic Mutation Finder (SMUFIN) (19) detects somatic SVs by comparing tumor and normal sequences without alignment to the reference sequence. This comparison requires a considerable amount of memory and computing time when it is applied to a whole-genome sequencing sample. For instance, SMUFIN requires more than 1 month using a 1.5 TB memory computer to detect SVs from whole-genome sequence data with 10x coverage (details in Supplementary Text). More efficient methods are thus highly desirable.

In this study, we introduce a precise, sensitive and computationally efficient somatic SV detection method, named COntrol SaMple-based detectiOn of Structural variation (COSMOS). COSMOS compares the mapping read status of paired-end short reads in a tumor sample with a normal sample in an asymmetric manner: groups of discordant read pairs, which are indicative of SVs, are generated

*To whom correspondence should be addressed. Tel: +81 3 3599 8915; Fax: +81 3 3599 8081; Email: sese.jun@aist.go.jp

from the tumor sample, following which the groups are filtered against individual discordant read pairs, instead of the group equivalents, in the normal sample to eliminate false positives. Next, we introduce the concept of strand-specific read depth, which allows prioritization of candidate SVs more efficiently than the conventional strand-independent read depth. Owing to these two unique properties, COSMOS outperforms other existing methods on synthetic as well as real data sets. In polymerase chain reaction (PCR)-based experiments, we confirmed that 84.5% of the SVs detected from mouse embryonic stem cells (ESCs) were correct, whereas the precision of the other methods were at most 70.4%. Moreover, our experimental results indicate that the sensitivity of COSMOS is comparable to the best alternative method.

MATERIALS AND METHODS

The COSMOS algorithm

COSMOS compares the statistics of paired-end reads in a tumor sample with a normal sample to detect SVs by incorporating two unique strategies: asymmetric comparison of the tumor sample versus the normal sample and a strand-specific read depth.

Figures 1 and 2 illustrate the procedure of COSMOS (Details in Supplementary Text, Supplementary Figures S1 and S2). Reads are obtained from one tumor sample and at least one normal sample, with a reference genome sequence also available. Reads from the tumor and normal samples are mapped on the reference genome independently. Our method can use any mapping program for the high-throughput sequencing reads, such as BWA (27) and Bowtie2(28).

COSMOS first divides read pairs into two categories: concordant and discordant. Concordant read pairs have span sizes (the distances between read pairs mapped on the reference genome) within the range of expected fragment sizes and consistent orientations of the read pairs with respect to the reference genome (see Supplementary Text for a formal definition). Discordant read pairs have unexpected span sizes and/or inconsistent orientations.

COSMOS then groups discordant read pairs whose genomic positions and span sizes are close to each other from the tumor sample (Figure 1B). Each group is considered to represent a different candidate SV. It is worth noting that, in this step, discordant read pairs having different span sizes are classified into discrete groups, even if their regions overlap, indicating that they most likely come from different SVs. Hence, COSMOS has the ability to detect SVs even when they overlap.

The candidate SVs represented by the discordant read-pair groups may include many false positives generated by misalignments attributed to repetitive sequences and/or allelic differences. To eliminate false positives, COSMOS compares the groups from the tumor sample with individual discordant read pairs from the normal sample instead of with the group equivalents from the normal samples (Figure 1C). Importantly, across the samples, the probability of accidental co-appearance of equivalent discordant read pairs is minimal. Therefore, the specificity of a group to the tumor sample is undercut by detecting even a single equivalent read pair in the normal sample. This asymmetric comparison procedure can remove many false positive groups in the tumor data, while maintaining sensitivity.

In the next step, COSMOS statistically selects SVs with high confidence values from the remaining groups using strand-specific read depth information (Figure 2 and Supplementary Figure S3). Even after removing the groups with an equivalent discordance in the normal samples, some false positives might remain, especially in the case of low read depth. In this situation, we propose using the strand-specific read depth. Read depth is usually calculated from the mapped reads without considering mapped strand specificity. However, COSMOS computes only the depths of concordant reads (Figure 2B), of which mapping orientations are either right-to-left (minus strand) or left-to-right (plus strand) (Figure 2C describes deletions and Supplementary Figure S3 describes other SV types). This way of computing read depth exhibits clearer depth borders than when combining information from both strands. For example, when we focus on concordant read pairs, the immediate upstream region of the deletion would contain only right-to-left directed reads because left-to-right directed reads would have unexpectedly large span sizes. Meanwhile, only left-to-right directed reads are produced from the immediate downstream region of the deletion. Within the deleted interval, no reads are detected. From these observations, a binomial distribution of the comparison in read depths between the inside and outside regions of each candidate SV gives the confidence score (Figure 2D). The candidate SVs are regarded as reliable SVs when their scores are larger than the lower range of the confidence interval. A larger score means a more reliable SV.

Implementation of COSMOS

We implemented COSMOS in open source Python software that is capable of detecting SVs from BAM files. COSMOS can easily be extended to handle multiple control samples, including the detection of *de novo* SVs in a family trio (proband, mother and father) by using the parents as controls.

Comparison with other methods

We compared COSMOS's performance with those of four widely used SV discovery software packages: BreakDancer(BD) (15), GASVPro(GASV) (29), DELLY (17) and LUMPY (18). BD is one of the most widely used SV detection software packages, and mainly uses the span size of paired-end reads like COSMOS. The others combine the span size information with split read alignment results. GASV and DELLY are selected due to their widespread use; LUMPY is a recently published method, whose accuracy is superior to that of GASV and DELLY on simulation data sets (18). LUMPY can detect SVs from multiple samples simultaneously.

We did not compare COSMOS's performance using human whole genome sequence data with that of SMUFIN (19), because SMUFIN required more than 1 month to detect SVs from the data. Moreover, tests on synthetic 30x reads from human chromosome 22 showed that COSMOS outperformed SMUFIN (Details in Supplementary Text).

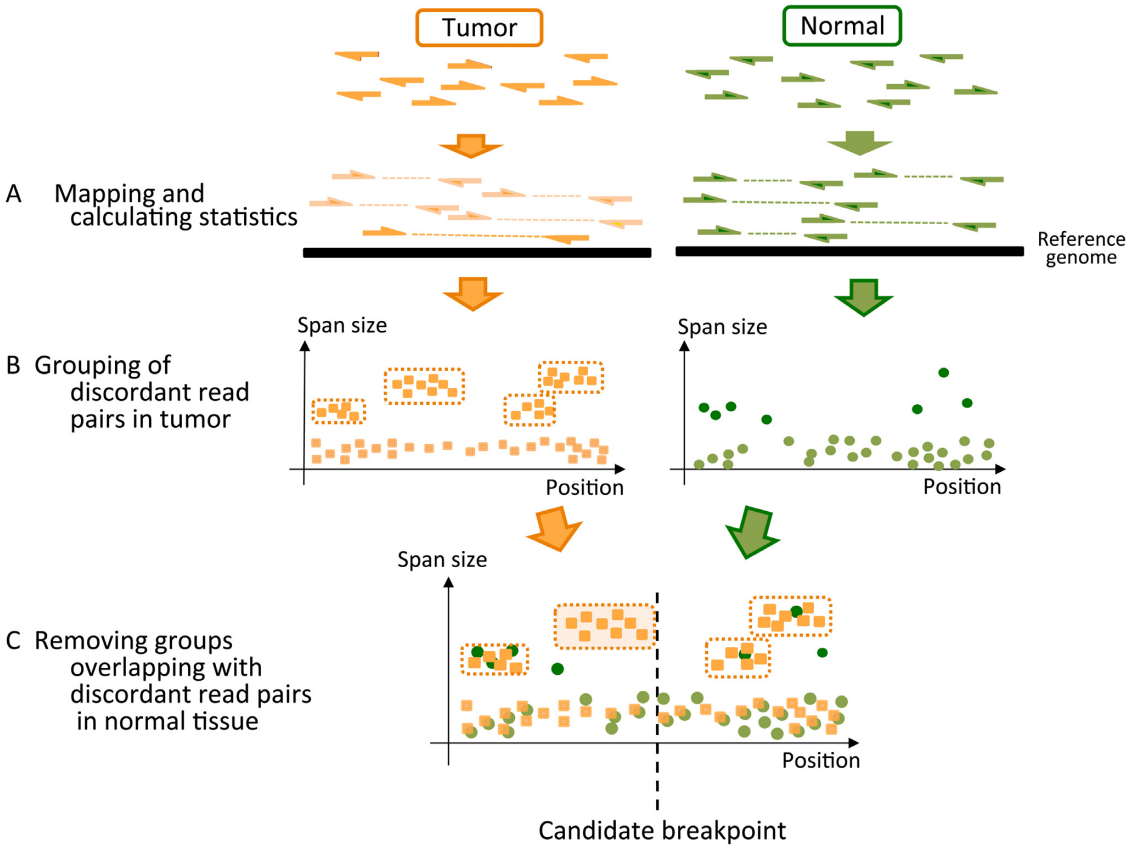

Figure 1. An asymmetric comparison in COSMOS. (**A**) Paired-end reads of tumor (orange) and normal (green) samples are independently mapped on the same reference genome. (**B**) According to the span sizes and chromosomal positions, groups of discordant read pairs are generated from the tumor sample. Each dot represents a left-to-right read in a mapped read pair. X- and Y-axes indicate the chromosomal position of each read and the span size of each pair, respectively. Each group consists of discordant reads with similar mapping positions and similar span sizes. Groups of similar mapping positions but different span sizes could be distinguished from each other. For the normal sample, COSMOS does not create groups (see text). (**C**) Groups whose corresponding areas on the XY plane overlap with any dot from the normal sample were removed from candidate groups. Rightmost positions in the remaining groups represent possible breakpoints of SVs at the nucleotide level.

Somatic mutation simulation data

We assessed the impact of read coverage, average insert size and standard deviation (SD) of insert size. The 'normal' genome was generated by introducing 8000 SVs (containing 2000 deletions, 2000 inversions, 2000 translocations and 2000 duplications) in the human reference sequence hg19. The 'tumor' sample contains two haploid genomes: one identical to the 'normal' genome, while the other was generated by introducing 800 additional SVs (200 deletions, 200 inversions, 200 translocations and 200 duplications) into the 'normal' genome. All SVs, except translocations, were introduced randomly throughout the genome with the span size ranging from 500 bp to 5 kbp. Translocations were introduced randomly, in a reciprocal fashion, between two different chromosomes. We used an in-house script to generate the SVs (available from the COSMOS website). We then used the WGSIM read simulator (https://github.com/lh3/wgsim) to sequence each simulated genome at 5x, 10x, 20x and 30x haploid coverage.

Mixture of tumor and normal samples

One of the difficulties in SV detection from tumor samples is that tumor tissue can be heterogeneous within the same lesion. We do not know, in advance, what percentage of the cells contain the tumor-specific genomes. To check the accuracies in this situation, we simulated the heterogeneous samples and varied the tumor genome content between 5% and 50%. In our results, COSMOS detected SVs with a F-measure of >0.9 when the tumor genome content was 10%.

Overlapping SV samples

We generated the rearranged genome as follows: first we generated a 'normal' diploid genome using the same method described in the previous subsection. We then generated a complex 'tumor' sample by randomly introducing 800 SVs into the paternal allele and 800 additional SVs to the maternal allele, so that the starting positions of the maternal SVs were identical to the parental ones, but the end positions were +1000 bp away from their parental counterparts.

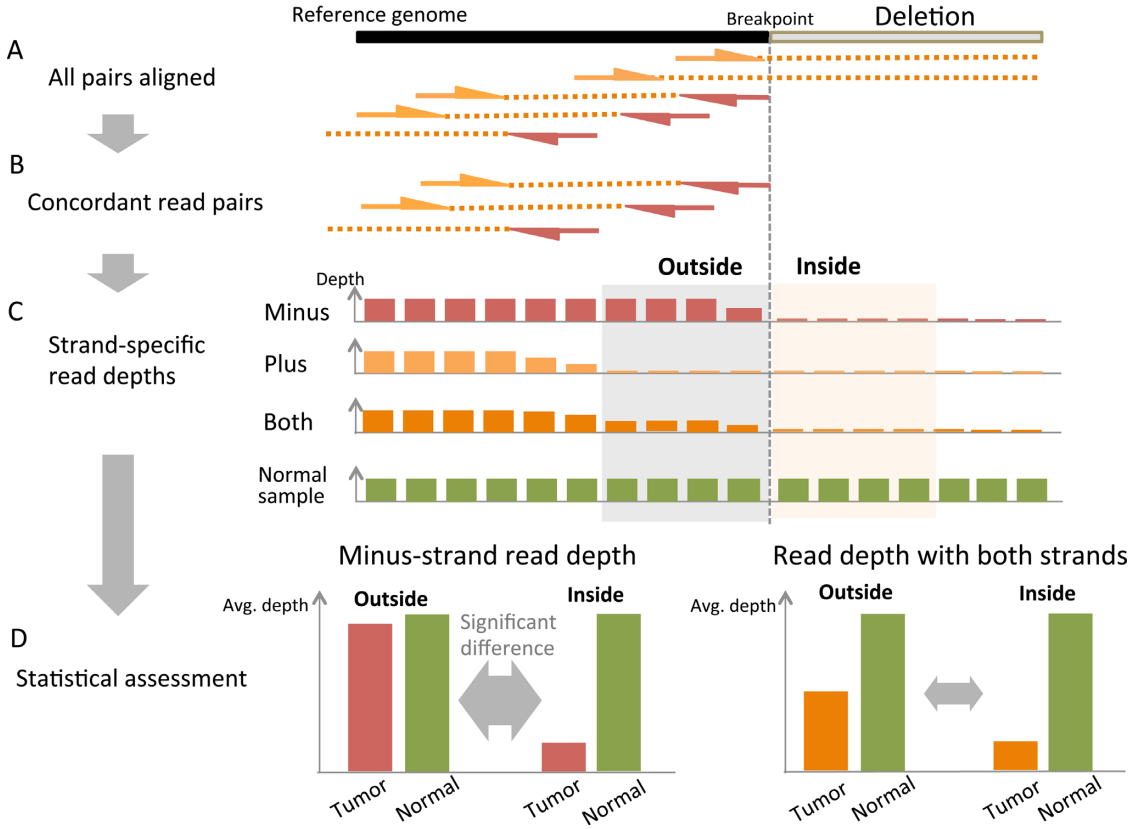

Figure 2. A statistical score of a SV based on strand-specific read depths. COSMOS computes statistical scores from the depths of strand-specific reads, providing more distinct differences in depth between the inside and outside regions of an SV, than with both strand-combined read depths. (**A**) Paired-end reads are aligned to the reference genome. Arrows represent positions and strand directions of the aligned reads: left-to-right (plus) and right-to-left (minus) strands of the reference (orange and red, respectively). Horizontal dashed lines (orange) represent sequencing gaps between the read pairs. (**B**) Only concordant read pairs are extracted. (**C**) In the immediate upstream region of, for example, a deletion-type SV, the depth of minus-strand reads is greater than that of plus-strand reads, which are mostly eliminated because of the unmappability of their opposite ends. As a control, the almost constant depth (green) of either strand reads in the normal sample is shown. Other types of SVs are described in Supplementary Figure S3. (**D**) Schematic of read-depth differences between the inside and outside regions of the deletion. The read depths are calculated in strand-dependent (left panel) and strand-independent (right panel) manners. To evaluate the read-depth differences, a statistical index is given by computing the confidence interval of the odds ratio between the inside and outside regions of the deletion; the higher the index, the more likely the SV exists.

Mouse embryonic stem cells

We next applied the SV detection methods to real short-read data obtained from a mouse ESC-based experimental 'tumor' model. To enhance the frequency of SVs, we used a genetically-engineered mouse ESC line ($Blm^{tet/tet}$), in which the expression of a genome-caretaker gene Blm can be transiently switched off by doxycycline (Dox) administration (31). The $Blm^{tet/tet}$ ESC line is derived from F1 hybrid mice between two inbred strains C57BL/6 (B6) and 129S4/SvJae (129), whose fully sequenced reference genome and single nucleotide polymorphism (SNP) information, respectively, are available (32). We introduced multiple SVs into the $Blm^{tet/tet}$ ESCs, by applying 8 Gy of gamma-irradiation, following destabilization of the cellular genome by transient Blm suppression. The cells were cultured to form colonies, which were then isolated and processed for Illumina's 90-bp paired-end whole-genome sequencing, with 30x coverage. As a 'normal' control, the isogenic ESCs, without Blm suppression or irradiation, were used. The paired-end reads were mapped to the B6 mouse reference genome (GRCm38/mm10) using Bowtie2, revealing an average insert size of 470 bp with an SD of 30 bp (Supplementary Figure S4).

RESULTS AND DISCUSSION

Performance comparison using simulation data

Our comparison showed that COSMOS achieves higher accuracy than the other methods across nearly all coverage and SV types (Figure 3A and Supplementary Figure S5). To compare the accuracies of the methods, we used the F-measure, calculated from the harmonic mean of precision and sensitivity as follows: 2* (precision * sensitivity) / (precision + sensitivity). This method was chosen as high sensitivity often causes many false positives, whereas low false-positive methods frequently generate conservative results with low sensitivity. The F-measure ranges from 0 to 1, and a high value means higher sensitivity and a lower false-positive rate at the same time. We checked the accuracies of the methods by varying the depth of reads. Figure 3A shows that BD, LUMPY and COSMOS achieved high F-measures for all SV types in the high coverage samples while GASV and DELLY scored low F-measures for at least one

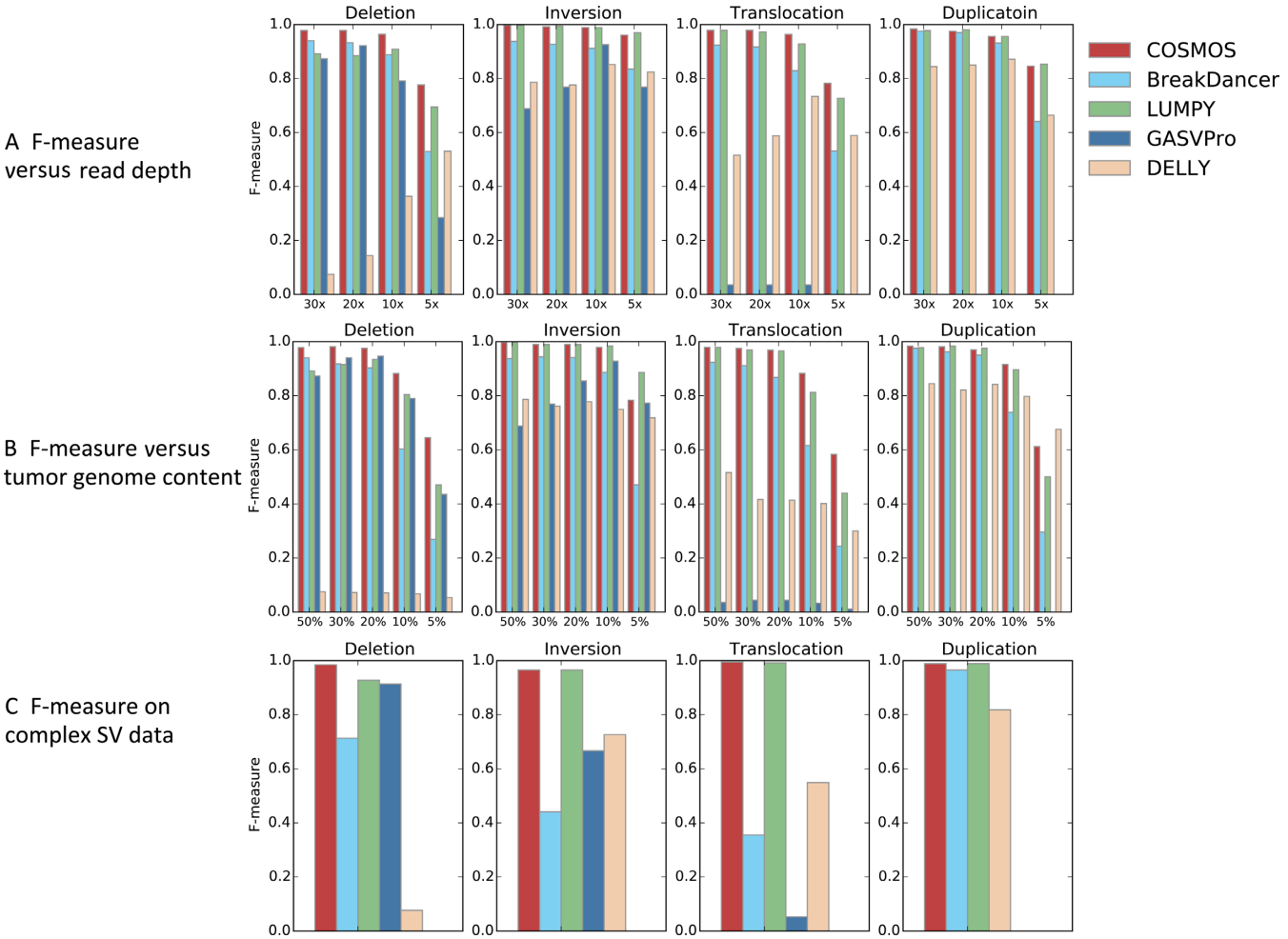

Figure 3. Performance comparison of COSMOS with other existing methods on simulation data sets. To measure the overall performance of each method in detecting deletion-type, inversion-type, translocation-type and duplication-type SVs. F-measure, a harmonic mean of sensitivity and precision, was computed on the different read depths (**A**), different tumor genome contents (**B**), different insert sizes (Supplementary Figure S6A) and different standard deviations of different insert sizes (Supplementary Figure S6B). (**C**) F-measure was computed for each method on a simulation data set, where mutually-overlapping SVs were artificially introduced. This data set is reminiscent of overlapping SVs, such as chromothripsis, in real tumors.

SV type. Indeed, GASV was conservative in detecting SVs, with few false positives (high precision), but with low sensitivity (Supplementary Figure S5A). By contrast, DELLY detected more SVs than necessary, causing numerous false positives that require costly and time-consuming confirmatory testing.

Of the methods that had high F-measures (BD, LUMPY and COSMOS), COSMOS had the highest score at most of the depths and SV types. COSMOS's scores were especially higher in deletion detection and in lower coverage samples. In the analysis of low-depth samples, strand-specific read depth served as a key determinant for the high performance of COSMOS. This is represented in Supplementary Figure S5, which shows that COSMOS, BD and LUMPY had similar sensitivities at all depths; nevertheless, BD precisions decreased dramatically as the depth decreased, and COSMOS obtained a higher precision than LUMPY, indicating that the comparison of normal and tumor samples (Figure 1C) performs well even if the normal sample is sequenced at a low coverage.

We next varied the average and SD of the insert sizes of the reads because COSMOS depends on this information to detect SVs. Supplementary Figure S6A shows that the F-measures of COSMOS are comparable to those of LUMPY even if the SD is large, indicating high tolerance to change in the SD of COSMOS. When the insert size was set to 200 bp, both the sensitivity and precision of the F-measures of COSMOS decreased (Supplementary Figure S6B) because many read pairs had no insert sequences. These results show that COSMOS has higher sensitivity and precision than the other methods for most paired-end libraries with >300 bp insert size.

We then compared the performances on mixture samples. Figure 3B shows that COSMOS's detection accuracy was comparable to that of LUMPY for a high tumor cell content and scored a higher F-measure than the other methods, especially when the content was low. Among the five methods, COSMOS, BD and LUMPY had higher F-measures than GASV and DELLY. Moreover, COSMOS was superior at deletion detection. Supplementary Figure S7 shows that all

of the methods exhibited decreased sensitivities as the tumor cell content decreased. However, the precision of COSMOS remained at 1.0, even when the tumor genome content was only 5% of the reads (Supplementary Figure S7A). DELLY had the highest sensitivity in all cases, while it had the lowest precision, thus indicating that DELLY yielded many false positives. The comparison between COSMOS and LUMPY indicated that, although their performances were similar in high tumor-content samples, COSMOS was superior to LUMPY in low tumor-content sample for detecting SVs except for sequence inversions.

Tumor genomes are often heterogeneous. Moreover, several recent genome analyses have shown that the SVs do not occur in random regions, but in positions close to each other (8,9,26,30). This means that the simulations described in the previous sections might be too simple. For a more complex case, we simulated two tumors whose regions of chromosomal rearrangements overlap, and compared the accuracies of COSMOS with the existing methods.

Figure 3C shows that COSMOS scored a higher F-measure compared with the other methods, even on the genome with overlapping SVs. Comparing COSMOS with LUMPY (having the second highest F-measure), we can see that only 1.02% of the deletions detected by COSMOS were false positives, while LUMPY was 11.1% (Supplementary Figure S8). Thus, COSMOS outperformed the other methods in this type of overlapping SV situation.

The consideration of span sizes when generating groups of discordant read pairs resulted in the high performance of COSMOS for detecting the overlapping SVs. When making these groups, sets of similar mapping positions but different span sizes were distinguished from each other (Figure 1B). Therefore, the two overlapping SVs were independently considered in the subsequent statistical assessment, and could be identified as independent SVs in COSMOS.

Statistical assessment with strand-specific read depth in COSMOS is efficient even if the estimated SV positions are slightly shifted from the real positions. Since COSMOS does not use split-read alignment, the estimated breakpoints might be slightly different from the real positions. We artificially shifted the breakpoints used in Figure 3A in a downstream direction, and measured true positive rates of the signal detection (Supplementary Figure S9). The true positive rates remains high even when the difference is 100 bp. Since the average difference between the estimated positions in COSMOS and the true break points is 14.7 bp (SD: 22.7 bp), the position offset does not affect the signal detection power in COSMOS.

COSMOS can identify SVs from human synthetic 30x reads in 6 h, while requiring less than 1 GB of memory with a single processor, which is comparable to other widely used methods for finding SVs.

SV detection in mouse embryonic stem cells

We next applied the SV detection methods to real short-read data obtained from a mouse ESC-based experimental 'tumor' model: (i) it detected 54 117 groups in total; (ii) among them, 825 SVs were selected by the 'asymmetric' comparison process between the tumor and normal samples; and (iii) binomial test scoring based on the strand-specific read depth, narrowed down the candidates to 84 SVs, which consisted of 31 deletions, 29 inversions, 12 translocations and 12 duplications (Supplementary Table S1). The minimum detected SV size was 75 bp. Notably, our subsequent subclonal analysis using array-comparative genomic hybridization (array-CGH) revealed that the tumor sample was a mixture of multiple (more than 4) subclones that harbor heterogeneous genomic rearrangements diverging during continuous cell proliferation. The process and mechanism of this genomic heterogeneity is currently being investigated as part of a follow-up study.

We compared the COSMOS results with those of BD and LUMPY, which achieved high F-measures in the previous simulation data analysis. When limited to $\geq$75 bp SVs and translocations, BD and LUMPY detected 1575 and 271 SVs, respectively (Figure 4A and Supplementary Tables S1 and S2). Those numbers were ~18.75- and 3.23-fold larger, respectively, than that obtained by COSMOS. Figure 4A shows that BD results did not overlap well with LUMPY and COSMOS. In contrast, almost all SVs detected by COSMOS were included in those found by LUMPY.

To estimate the sensitivity and the precision of the results, we arbitrarily selected 58 out of 84 SV candidates predicted by COSMOS for PCR-based experimental validation (Figure 4B and Supplementary Tables S1 and S3). As a result, 49 out of the 58 SV candidates (84.5%) were successfully confirmed by junction PCR and Sanger sequencing. Similarly, we checked 71 out of 271 SV candidates predicted by LUMPY, and confirmed 50 of them (70.4%) by PCR-Sanger sequencing. These experimentally confirmed SVs include the 49 SVs predicted by COSMOS, indicating that the vast majority of COSMOS and LUMPY double-positive SVs are true positives. The average differences of the estimated break point positions in COSMOS from the detected positions using Sanger sequences were comparable to those in LUMPY (Supplementary Figure S10). Conversely, 17 SVs, which were detected by LUMPY but not by COSMOS, were likewise subjected to PCR-based validation. Among the 17 SVs, only one (5.9%) was experimentally confirmed, indicating that the SV candidates, predicted as positive by LUMPY but as negative by COSMOS, contain a high rate of false positives.

We next focused on 33 SVs detected by both COSMOS and BD (Figure 4A). Twenty-eight of 33 SVs (84.8%) were confirmed by PCR and Sanger sequencing. BD overlooked at least 22 SVs, showing low sensitivity.

CONCLUSION

We developed a highly sensitive and accurate SV detection method using both the asymmetric comparison of tumor and normal samples and a confidence score statistically calculated from the strand-specific read depth. Our method outperformed existing SV detection methods, even when the tumor cell contents were relatively low in the sample tissue, and two different 'tumor' genomes were present in the sample tissue.

On real data sets of mouse whole-genome sequencing, we confirmed that the precision of COSMOS's prediction for SVs is 84.5%, which is at least 14.1% higher than that of other methods. Moreover, although it was difficult to detect

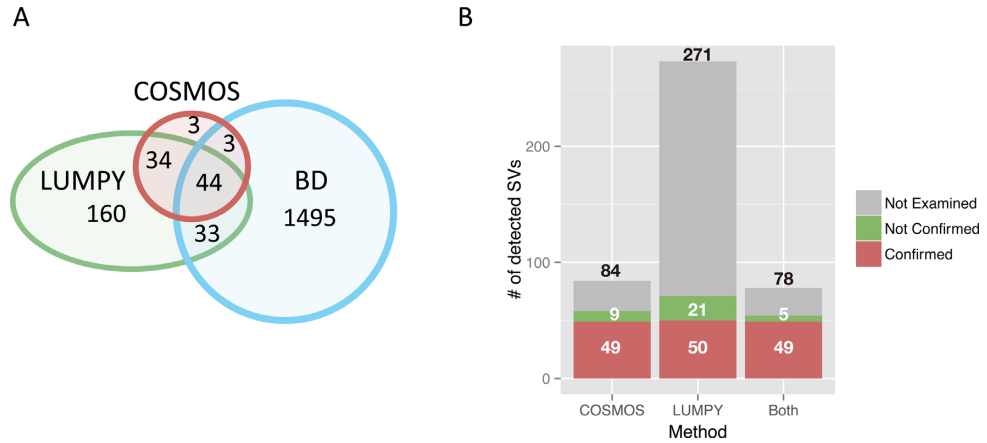

Figure 4. Performance comparison of COSMOS with other methods on real mouse ESC data sets of paired-end whole-genome sequencing data. (**A**) Venn diagram of SVs detected by COSMOS, BD and LUMPY. The number of SVs is shown in the corresponding category. (**B**) Results of PCR-based experimental validation of SVs detected by COSMOS, by LUMPY and by both methods.

all SVs in the cells, the PCR-based validation experiments demonstrated that the sensitivity of COSMOS was among the highest of the existing methods.

COSMOS focuses on the detection of relatively large SVs using the statistics of span sizes and strand-specific mapped reads. The combination of COSMOS with split-reads alignments (17,18,29), a reference-free strategy (19,33) and control-free approach (34) might allow us to detect smaller SVs, thereby yielding a more comprehensive detection of SVs.

The problem of detecting SVs by comparing a normal sample with a tumor sample has theoretical limitations such that unique solution cannot be determined when the SVs overlap. For example, given a genomic region A-B-C-D-E, an inversion of B-C generates an A-C'-B'-D-E genome, where B' and C' indicate the inverted sequences. Then, an inversion of B'-D generates an A-C'-D'-B-E genome. An SV detection on the genome may detect an inversion C'-D', an insertion between A and B, and a deletion between B and E, representing a different explanation of the SV generation process. In future work, we would like to determine the best possible SVs automatically, perhaps by introducing an assumption such as a minimum number of rearrangements (35).

ACCESSION NUMBERS

The accession numbers of the sequences used in this paper are SAMD00020209 and SAMD00020213.

AVAILABILITY

All programs to detect SVs are available from http://seselab.org/cosmos/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

Author contributions: K.Y. and J.S. contributed new analytic tools and performed computational analysis. A.Y. and C.K. performed experimentally validation of biological experiments. C.K., J.T. and J.S. designed the research. K.Y., C.K., J.T. and J.S. wrote the manuscript.

J.S. thanks Tony Kuo for helpful discussions.

FUNDING

The supercomputing resource was provided by the National Institute of Genetics, Research Organization of Information and Systems, Japan; Japan Society for the Promotion of Science (JSPS) KAKENHI [24240044, 25125709 and 15H01717 to J.S., 25125714 and 221S0002 to J.T.]; Grant-in-Aid for Scientific Research on InnovativeArea "Genome Science", Ministry of Education, Culture, Sports, Science and Technology in Japan [to J.S.]. Funding for open access charge: AIST.

Conflict of interest statement. None declared.

REFERENCES

1. Yates,L.R. and Campbell,P.J. (2012) Evolution of the cancer genome. *Nat. Rev. Genet.*, **13**, 795–806.
2. Yang,L., Luquette,L.J., Gehlenborg,N., Xi,R., Haseley,P.S., Hsieh,C.-H., Zhang,C., Ren,X., Protopopov,A., Chin,L. *et al.* (2013) Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell*, **153**, 919–929.
3. Zhang,F., Gu,W., Hurles,M.E. and Lupski,J.R. (2009) Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.*, **10**, 451–481.
4. Marshall,C.R., Noor,A., Vincent,J.B., Lionel,A.C., Feuk,L., Skaug,J., Shago,M., Moessner,R., Pinto,D., Ren,Y. *et al.* (2008) Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.*, **82**, 477–488.
5. Weischenfeldt,J., Symmons,O., Spitz,F. and Korbel,J.O. (2013) Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.*, **14**, 125–138.
6. Ledford,H. (2010) Big science: the cancer genome challenge. *Nature*, **464**, 972–974.
7. Baker,M. (2012) Structural variation: the genome's hidden architecture. *Nat. Methods*, **9**, 133–137.
8. Zhang,C.Z., Leibowitz,M.L. and Pellman,D. (2013) Chromothripsis and beyond: rapid genome evolution from complex chromosomal rearrangements. *Genes Dev.*, **27**, 2513–2530.
9. Stephens,P.J., Greenman,C.D., Fu,B., Yang,F. and Bignell,G.R. (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, **144**, 27–40.

10. Korbel,J.O. and Campbell,P.J. (2013) Criteria for Inference of chromothripsis in cancer genomes. *Cell*, **152**, 1226–1236.
11. Quinlan,A.R. and Hall,I.M. (2012) Characterizing complex structural variation in germline and somatic genomes. *Trends Genet.*, **28**, 43–53.
12. Zhang,F., Carvalho,C.M.B. and Lupski,J.R. (2009) Complex human chromosomal and genomic rearrangements. *Trends Genet.*, **25**, 298–307.
13. Pabinger,S., Dander,A., Fischer,M., Snajder,R., Sperk,M., Efremova,M., Krabichler,B., Speicher,M.R., Zschocke,J. and Trajanoski,Z. (2014) A survey of tools for variant analysis of next-generation genome sequencing data. *Brief. Bioinform.*, **15**, 256–278.
14. Ding,L., Wendl,M.C., McMichael,J.F. and Raphael,B.J. (2014) Expanding the computational toolbox for mining cancer genomes. *Nat. Rev. Genet.*, **15**, 556–570.
15. Chen,K., Wallis,J.W., McLellan,M.D., Larson,D.E., Kalicki,J.M., Pohl,C.S., McGrath,S.D., Wendl,M.C., Zhang,Q., Locke,D.P. *et al.* (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods*, **6**, 677–681.
16. Ye,K., Schulz,M.H., Long,Q., Apweiler,R. and Ning,Z. (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**, 2865–2871.
17. Rausch,T., Zichner,T., Schlattl,A., Stutz,A.M., Benes,V. and Korbel,J.O. (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**, i333–i339.
18. Layer,R.M., Chiang,C., Quinlan,A.R. and Hall,I.M. (2014) LUMPY: A probabilistic framework for structural variant discovery. *Genome Biol.*, **15**, R84.
19. Moncunill,V., Gonzalez,S., Beà,S., Andrieux,L.O., Salaverria,I., Royo,C., Martinez,L., Puiggròs,M., Segura-Wang,M., Stütz,A.M. *et al.* (2014) Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nat. Biotechnol.*, **32**, 1106–1112.
20. Rizk,G., Gouin,A., Chikhi,R. and Lemaitre,C. (2014) MindTheGap : integrated detection and assembly of short and long insertions. *Bioinformatics*, **30**, 3451–3457.
21. Abyzov,A., Li,S., Kim,D.R., Mohiyuddin,M., Stütz,A.M., Parrish,N.F., Mu,X.J., Clark,W., Chen,K., Hurles,M. *et al.* (2015) Analysis of deletion breakpoints from 1, 092 humans reveals details of mutation mechanisms. *Nat. Commun.*, **6**, 7256.
22. Wang,J., Mullighan,C.G., Easton,J., Roberts,S., Heatley,S.L., Ma,J., Rusch,M.C., Chen,K., Harris,C.C., Ding,L. *et al.* (2011) CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat. Methods*, **8**, 652–654.
23. Lam,H.Y.K., Mu,X.J., Stütz,A.M., Tanzer,A., Cayting,P.D., Snyder,M., Kim,P.M., Korbel,J.O. and Gerstein,M.B. (2009) Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat. Biotechnol.*, **28**, 47–55.
24. Marusyk,A. and Polyak,K. (2010) Tumor heterogeneity: causes and consequences. *Biochim. Biophys. Acta*, **1805**, 105–117.
25. Burrell,R.A., McGranahan,N., Bartek,J. and Swanton,C. (2013) The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, **501**, 338–345.
26. Chiang,C., Jacobsen,J.C., Ernst,C., Hanscom,C., Heilbut,A., Blumenthal,I., Mills,R.E., Kirby,A., Lindgren,A.M., Rudiger,S.R. *et al.* (2012) Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. *Nat. Genet.*, **44**, 390–397.
27. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
28. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
29. Sindi,S.S., Önal,S., Peng,L.C., Wu,H.-T. and Raphael,B.J. (2012) An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biol.*, **13**, R22.
30. De,S. and Michor,F. (2011) DNA secondary structures and epigenetic determinants of cancer genome evolution. *Nat. Struct. Mol. Biol.*, **18**, 950–955.
31. Yamanishi,A., Yusa,K., Horie,K., Tokunaga,M., Kusano,K., Kokubu,C. and Takeda,J. (2013) Enhancement of microhomology-mediated genomic rearrangements by transient loss of mouse Bloom syndrome helicase. *Genome Res.*, **23**, 1462–1473.
32. Keane,T.M., Goodstadt,L., Danecek,P., White,M.A., Wong,K., Yalcin,B., Heger,A., Agam,A., Slater,G., Goodson,M. *et al.* (2011) Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, **477**, 289–294.
33. Wijaya,E., Shimizu,K., Asai,K. and Hamada,M. (2014) Reference-free prediction of rearrangement breakpoint reads. *Bioinformatics*, **30**, 2559–2567.
34. Boeva,V., Zinovyev,A., Bleakley,K., Vert,J.-P., Janoueix-Lerosey,I., Delattre,O. and Barillot,E. (2011) Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics*, **27**, 268–269.
35. Pevzner,P. and Tesler,G. (2003) Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res.*, **13**, 37–45.