# BSSV: Bayesian based Somatic Structural Variation Identification with Whole Genome DNA-Seq Data

Xi Chen[1], Xu Shi[1], Ayesha N. Shajahan[2], Leena Hilakivi-Clarke[2], Robert Clarke[2] and Jianhua Xuan[1,*]

*Abstract*— **High coverage whole genome DNA-sequencing enables identification of somatic structural variation (SSV) more evident in paired tumor and normal samples. Recent studies show that simultaneous analysis of paired samples provides a better resolution of SSV detection than subtracting shared SVs. However, available tools can neither identify all types of SSVs nor provide any rank information regarding their somatic features. In this paper, we have developed a Bayesian framework, by integrating read alignment information from both tumor and normal samples, called BSSV, to calculate the significance of each SSV. Tested by simulated data, the precision of BSSV is comparable to that of available tools and the false negative rate is significantly lowered. We have also applied this approach to The Cancer Genome Atlas breast cancer data for SSV detection. Many known breast cancer specific mutated genes like RAD51, BRIP1, ER, PGR and PTPRD have been successfully identified.**

## I. Introduction

Somatic mutations, which drive cancer development, are acquired during a person's lifetime and cause tumor cells to divide faster than normal. Some mutations occur within the gene itself, while others at the promoter regions that control the transcription of genes. One major type of mutation is structural variation (SV), including deletion, insertions, inversions and translocations subtypes [1]. It is known that the fraction of the genome affected by SVs is comparatively larger than that accounted for by single nucleotide polymorphisms and other small scale variants [2]. Thus, the contribution of SVs to cancer related genetic variation analysis is becoming increasingly important.

Deep DNA-sequencing on whole genome has enabled the SV identification at base-pair resolution, providing precise genomic locations of breakpoints for most types of SVs. A typical approach is Breakdancer [3], which aligns the paired-end reads (PR), sequenced from test genome onto the reference genome and looks for 'discordant' PRs that may indicate the presence of SVs nearby. More recent methods, like GASVPro [4] and PeSV-Fisher [5], have integrated PR and read depth (RD) signals to increase the sensitivity of identifying a segment deletion. Additionally, Pindel [6] and Delly [7] incorporated splitting read (SR) signals to further improve the precision of breakpoints. In earlier works, somatic region extraction was achieved by identifying SVs from tumor and normal samples independently and subtracting the shared results [8]. Actually, each of the SV identification tools mentioned above is able to produce a list of non-shared variants in paired samples. However, a potential problem comes up as such an approach suffers a high risk to miss out on some somatic SVs (SSVs) due to the false positive predictions in normal samples. Some recent tools on small indel identification have developed generalized Bayesian framework to call somatic regions by comparing genomic changes in both samples simultaneously [9], but few algorithms have been developed for SSV detection. Seurat [9] has been developed for small somatic variants detection by assessing the similarity of genomic changes between tumor and normal samples from probabilistic aspect. The observations from either sample increase the detection sensitivity by providing more evidence towards a somatic change or a germline mutation.

To improve the quality of (SSV) identification, statistical methods analyzing both tumor and normal samples simultaneously are needed, which provide a measurement of confidence level for each candidate SSV. In this paper, we have developed a SSV identification method, under Bayesian framework, called BSSV, to calculate significance *p*-value for each SSV by comparing the read alignments in tumor sample to those observed in normal sample. Two major steps include extracting all SV regions from tumor sample and examining read alignments in both samples at each region. Rather than using a uniform model, we investigated in detail how each kind of SV is formed and what information can be used to determine its relative importance. For each SSV subtype, we developed a specific model in the proposed BSSV approach. We used simulated DNA-Seq data to demonstrate the efficiency of BSSV and further compared the precision/recall performance to several available tools [3-5] . Overall, BSSV has a comparable performance on the precision but provides a significantly higher recall value than that achieved by the available tools. Further, we applied BSSV to The Cancer Genome Atlas (TCGA) breast cancer patient samples. When compared to the results generated by Breakdancer, we identified more somatically mutated genes.

## II. Methods

A general step in SV identification is to cluster discordant reads into clusters [3, 4, 10]. The determination of discordant read is based on insert size distribution and alignment orientation between paired reads. A candidate SV is derived from a region that is interconnected by at least two discordant

*Corresponding author

This research was supported in part by NIH Grants (CA149653, CA139246, CA149147 and CA164384).

Xi Chen, Xu Shi and Jianhua Xuan are with the Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203 USA (e-mail: {xichen86, xushi, xuan}@vt.edu).

Ayesha N. Shajahan, Leena Hilakivi-Clarke, Robert Clarke are with Department of Oncology, School of Medicine, Georgetown University, Washington, DC 20057, USA; (e-mail: {ans33, clarkel, clarker}@georgetwon.edu).

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

reads and further assigned as a deleted fragment, an insert fragment, or an inverted region. To increase the sensitivity in SSV identification, we refined the SVs by setting the number of supporting discordant reads in tumor sample to at least four and also larger than the number of discordant reads in normal sample at the same genome location. We incorporated RD signal in the modeling part to improve the sensitivity of SSV detection. Instead of counting the number of concordant reads, we used GATK [11] to calculate average RD within and at both flanks of each mutation region.

### A. Somatic deletion identification

To estimate the confidence level of each candidate somatic mutation, we define $P_{somatic}(SV \mid s)$ as the significance *p*-value of a SSV being detected at $s$. To avoid the impact of coverage difference between tumor and normal samples, the number of reads at region $s$ in normal sample is normalized using RD ratio at flank regions between tumor and normal samples. In this way, the read counts in both samples, no matter discordant or concordant, are comparable. In general, the somatic feature is determined by comparing discordant read count $k_D$ and concordant read count $k_C$ at region $s$ in tumor sample ($T$) to the observations in normal sample ($N$) as follows:

$$P_{somatic}(SV \mid s) = P(k_{D,T}, k_{C,T} \mid k_{D,N}, k_{C,N}). \quad (1)$$

A direct observation of somatic deletion is that fewer (heterozygous) or almost no (homozygous) concordant reads from tumor sample are mapped to the reference genome compared to those mapped in normal sample to the same region. And a cluster of discordant reads with significantly longer insert size can be identified at breakpoint boundary region. To determine the *p*-value of a somatic deletion, compared to normal sample, we looked at the joint probability of discordant reads increase and concordant reads loss in tumor sample. As demonstrated in [4], these two components are independent because they are located at close but different genome locations (within and outside the variation boundary). Thus, (1) can be divided into two conditional terms as:

$$\begin{aligned} P_{somatic}(DEL \mid s) &= \frac{P(k_{D,T}, k_{D,N})P(k_{C,T}, k_{C,N})}{P(k_{D,N})P(k_{C,N})} \\ &= P(k_{D,T} \mid k_{D,N})P(k_{C,T} \mid k_{C,N}) \end{aligned}, \quad (2)$$

where $k_C$ represents the number of concordant reads covering the midpoint of the deletion region, and $k_D$ is the number of discordant reads at breakpoint.

$P_{somatic}(DEL \mid s)$ provides a measurement of read alignment difference between tumor and normal samples for a deletion region. The first term in (2) evaluates whether the discordant reads in tumor sample is significantly larger than those in normal sample. We can use Bayesian rule to equivalently calculate the probability as (3). This Bayesian transform is reasonable here since all candidate SSVs are first identified from tumor sample, and then checked in normal sample.

$$P(k_{D,T} \mid k_{D,N}) = P(k_{D,N} \mid k_{D,T})P(k_{D,T}) / P(k_{D,N}), \quad (3)$$

To calculate the first term in (3), we define a success as a discordant read observed in tumor sample, while a failure as a discordant read observed in normal sample. Then, we use negative binomial model to calculate the probability of observing $k_{D,N}$ failures in normal sample in a sequence of Bernoulli trials before $k_{D,T}$ successes occurring in tumor sample:

$$P(k_{D,N} \mid k_{D,T}) = \sum_{i=1}^{k_{D,N}} \binom{i + k_{D,T} - 1}{i} (1 - p_{del})^i p_{del}^{k_{D,T}}, \quad (4)$$

where success rate $p_{del}$ is defined as $K_{del,T} / (K_{del,T} + K_{del,N})$. $K_{del,T}$ and $K_{del,N}$ are the total numbers of deletion type discordant reads in tumor and normal samples, respectively.

The prior probability of observing $k_{D,T}$ and $k_{D,N}$ discordant reads in tumor and normal genome region $s$ can be calculated by using Poisson model [3] as (5):

$$\begin{cases} P(k_{D,T}) = \text{Poisson}_{pmf}(k_{D,T}, \lambda_{del,T}) \\ P(k_{D,N}) = \text{Poisson}_{pmf}(k_{D,N}, \lambda_{del,N}) \end{cases}, \quad (5)$$

where $\lambda_{del,T(N)}$ equals to $K_{del,T(N)} \cdot s / (0.5G)$, and $0.5G$ is the effective length of whole genome.

Concordant component $P(k_{C,T} \mid k_{C,N})$ in (2) represents the probability that $k_{C,N} - k_{C,T}$ reads in tumor sample are deleted out of $k_{C,N}$ concordant reads in normal sample. It can be calculated by using binomial cumulative distribution $B_{cmf}$, as follows:

$$P(k_{C,T} \mid k_{C,N}) = B_{cmf}(k_{C,T}, k_{C,N}, p_c), \; p_c = 0.5. \quad (6)$$

By bringing (3) ~ (6) to (2), we could calculate the *p*-value for somatic deletion. The lower this value is, the more confidence we obtain for this somatic change.

### B. Somatic insertion identification

For insertion variation, two ends of discordant reads are mapped very close because their covered genome region is either unknown or inserted from other places. In this case, only these closely mapped discordant reads in paired samples are used for somatic insertion significance evaluation.

$$P_{somatic}(INS \mid s) = P(k_{D,T} \mid k_{D,N}). \quad (7)$$

According to Bayesian rule, (7) can be extended into likelihood and prior format as:

$$P_{somatic}(INS \mid s) = P(k_{D,N} \mid k_{D,T})P(k_{D,T}) / P(k_{D,N}). \quad (8)$$

The calculation of each component is similar to (4) and (5). Here, the mean parameter in $\lambda_{ins,T(N)}$ Poisson model equals to $K_{ins,T(N)} \cdot (\mu - s) / (0.5G)$, where $K_{ins,T(N)}$ is the total number of insertion type reads in tumor and normal samples, respectively and $\mu$ is the mean of insert size distribution.

### C. *Somatic inversion identification*

The cause of a somatic inversion is that a batch of reads in the tumor sample is mapped to the reference genome with one side transposed due to segment inversion. Near the boundary region of inversion, the more we observe the discordant reads, the less we can map the concordant reads to the reference genome. Hence, discordant reads $k_D$ and concordant reads $k_C$ are related at breakpoints. And after normalization, the discordant reads in both samples are directly comparable. Here, the length of inverted segment is not important because it doesn't impact the number of reads around breakpoints. Therefore, for somatic inversion, discordant reads with one end orientation change are used to evaluate the significance. According to Bayesian rule, (1) can be expressed as:

$$P_{somatic}(INV \mid s) = P(k_{D,T} \mid k_{D,N}) = P(k_{D,N} \mid k_{D,T}) \frac{P(k_{D,T})}{P(k_{D,N})}. \quad (9)$$

Similar to (4), negative binomial model is used to calculate the likelihood term $P(k_{D,N} \mid k_{D,T})$. Considering the dependency between discordant and concordant reads at breakpoints of segment inversion, the prior probabilities $P(k_{D,T})$ and $P(k_{D,N})$ in (9) are calculated by using binomial model as:

$$\begin{cases} P(k_{D,T}) = B_{pmf}(k_{D,T}, C, p_{inv,T}) \\ P(k_{D,N}) = B_{pmf}(k_{D,N}, C, p_{inv,N}) \end{cases}, \quad (10)$$

where $C$ is the normalized read coverage at breakpoint, as a sum of discordant and concordant reads, which is the same for both samples. The success rate $p_{inv,T(N)}$ is defined as $K_{inv,T(N)} / K_{C,T(N)}$. $K_{inv,T(N)}$ and $K_{C,T(N)}$ are the total numbers of inversion type discordant reads and concordant reads in tumor and normal samples.

As mentioned above, the formation of each subtype of SSV is different. To identify SSVs more accurately, we developed individual model in our BSSV approach for each subtype of SSV, including deletion, insertion and inversion. Together they provide a complete picture of SSV identification in this study.

## III. RESULTS

We tested our proposed BSSV approachon both simulated and real data sets to demonstrate its superiority over existing methods for SSV detection.

### A. *Evaluation of BSSV with simulated data*

First, to evaluate our BSSV's sensitivity and specificity on somatic SV detection, we simulated a pair of genomes as 'tumor' and 'normal' by using RSVSIM [12]. For germline mutations, we generated 50 variants each for deletion, insertion and inversion on chromosome 22 extracted from human genome (hg19 ). The length of deletion and inversion follows uniform distribution varying from 500 to 1500 bps. The size of insertion is from 50 to 200 bps. For each subtype, another 50 somatic variants are added to the 'tumor' genome only. WGSIM [13] is then used to sequence each genome with following settings: Illumina platform, read length 100 bps, insert size mean 300 bps, standard deviation 30 bps, and average RD as 10, 20 and 40, respectively. The average RD is the most important factor affecting the sensitivity of mutation detection [5]. We designed different scenarios to test our proposed method. Sequenced data for both genomes are first aligned to reference genome with BWA [14]. Discordant reads clustering and somatic region detection are then conducted using our proposed model.

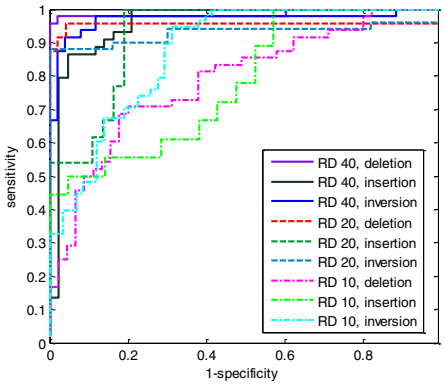


Figure 1. ROC performance for somatic deletion, insertion and invertion identifictaion with BSSV.

The receiver operating characteristic (ROC) curve for the detection of each subtype of SSV is shown in Fig. 1. When the average RD is 40, the ROC performance is high with area under the curve (AUC) as 0.96, 0.94, and 0.95 for deletion, insertion and inversion, respectively. When the RD is at medium level, for deletion and inversion, the AUC values are 0.934 and 0.908. We missed 2 somatic deletions and another 2 somatic inversions during the discordant reads clustering step. The sensitivity can only reach up to 0.96. While for insertion, which is a more challenging case, the AUC value drops faster to 0.88. The detection performance is lower because fewer discordant reads can be mapped successfully around insertion region. The average RD for real data is usually between 20 to 40 folds. Within this range, our BSSV could provide a steady performance on SSV detection. We also tested our proposed method on a larger data set with 100 somatic variants for each type. The AUC values for deletion and inversion are 0.97 and 0.92, respectively. The impact of few missed regions is lowered. For insertion, the AUC value is 0.86. Since BSSV tests candidate mutation regions one by one, the number of mutations on the genome doesn't affect the detection performance much.

### B. *Comparison with available tools*

We further compared the proposed BSSV to several published tools by using simulated data with medium RD 20. Table I shows their precision/recall performance.

TABLE I. PRECISION/RECALL PERFORMANCE FOR SOMATIC CALLING

| SV | BSSV | | Breakdancer | | GASVPro | | PeSV-Fisher | |
|---|---|---|---|---|---|---|---|---|
| | *Pre./Recall* | | *Pre./Recall* | | *Pre./Recall* | | *Pre./Recall* | |
| DEL | 0.94 | 0.94 | 0.75 | 0.90 | 0.74 | 0.86 | 0.94 | 0.60 |
| INS | 0.82 | 0.82 | 0.74 | 0.74 | - | - | - | - |
| INV | 0.94 | 0.94 | 0.92 | 0.92 | 0.73 | 0.60 | 0.98 | 0.60 |

It can be noted that for somatic deletion detection, BSSV and PeSV-Fisher achieve the highest precision at 0.94. But PeSV-Fisher has a very low recall at 0.6 due to directly deleting common SVs for paired samples after discordant read clustering. Breakdancer and GASV capture most SSVs with recall around 0.9 but both have a high false positive rate around 0.25. For insertion, among three published approaches, only Breakdancer identifies a batch of candidate regions. Our method gives a higher performance if compared to Breakdancer. Inversion type of SV is easier to be detected by checking read orientation. Even though BSSV and Breakdancer achieve similar performance, BSSV provides a ranked list of inversions rather than treating each SSV equally. Their recall values are significantly better than other tools, reporting more true somatic inversions. Overall, BSSV works better on the SSV detection than several widely used tools.

### C. TCGA breast cancer SSV detection

To detect SSVs in real study, we applied the proposed BSSV to whole genome DNA-Seq data of 14 TCGA ER-negative breast cancer patients with chemotherapy (https://tcga-data.nci.nih.gov/tcga/). Their survival time has a mean value of 2 years with a variance of 1.3 years, and their tumor cells are quite aggressive. We processed each pair of tumor/normal samples with BSSV independently and selected the SSVs with $p$-value < 1e-6 for each patient. We also applied Breakdancer to this data set to compare results. As claimed in [15], most of the somatic mutations occurred at less than 10% incidence across all breast cancers. To lower the false positive rate, with 14 samples, we selected SSVs shared by at least two patients. As shown in Fig.2, BSSV reports more deletions and insertions. This is consistent to our hypothesis that subtracting SV lists from paired samples, whereas Breakdancer would miss some SSVs. In simulation study, BSSV and Breakdancer have quite similar performance on inversion detection. For real data, their results are also quite similar, but BSSV provides confidence information for each reported SSV.

By comparing to genes previously implicated in breast cancer [15], we observed somatic deletions at CCND3, PTPRD, MLL3, and RAD51, an insertion at BRIP1 and inversions at AFF2, ESR1, PGR, NF1, PIK3CA, PTPRD, and RB1. It is not strange that breast cancer specific mutations like ERS1 and PGR are observed. PIK3CA is an oncogene showing the highest frequency of gain-of-function mutations in breast cancer. PTPRD is a tumor suppresser, the somatic deletions on which will alleviate growth suppression and apoptosis . In addition, genomic changes in RAD51 and BRIP1 likely disrupt the normal DNA repair process, and allow the cells to grow and divide uncontrollably eventually forming a tumor.

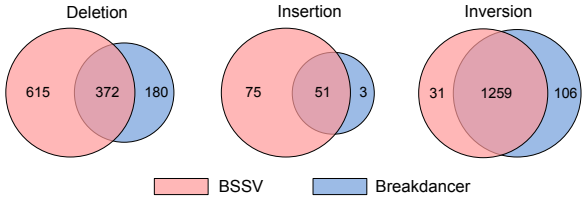


Figure 2. Detected SSVs of TCGA breast cancer patient DNA-Seq data.

## IV. Conclusion

We presented a Bayesian based SSV detection method, BSSV, to identify cancer specific genomic changes. BSSV processed tumor and normal samples simultaneously and identified more accurate SSVs than conventional tools. Significance $p$-value calculated by BSSV for each SSV can provide biologists a ranked list for further experimental validation. A practical extension to current approach is to simultaneously integrate read information from multiple samples to lower the false positive rate in identified SSVs.

## References


[1] P. Medvedev, M. Stanciu, and M. Brudno, "Computational methods for discovering structural variation with next-generation sequencing," *Nat Methods,* vol. 6, pp. S13-20, Nov 2009.

[2] D. F. Conrad, D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, *et al.*, "Origins and functional impact of copy number variation in the human genome," *Nature,* vol. 464, pp. 704-12, Apr 1 2010.

[3] K. Chen, J. W. Wallis, M. D. McLellan, D. E. Larson, J. M. Kalicki, C. S. Pohl, *et al.*, "BreakDancer: an algorithm for high-resolution mapping of genomic structural variation," *Nat Methods,* vol. 6, pp. 677-81, Sep 2009.

[4] S. S. Sindi, S. Onal, L. C. Peng, H. T. Wu, and B. J. Raphael, "An integrative probabilistic model for identification of structural variation in sequencing data," *Genome Biol,* vol. 13, p. R22, 2012.

[5] G. Escaramis, C. Tornador, L. Bassaganyas, R. Rabionet, J. M. Tubio, A. Martinez-Fundichely, *et al.*, "PeSV-Fisher: identification of somatic and non-somatic structural variants using next generation sequencing data," *PLoS One,* vol. 8, p. e63377, 2013.

[6] K. Ye, M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning, "Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads," *Bioinformatics,* vol. 25, pp. 2865-71, Nov 1 2009.

[7] T. Rausch, T. Zichner, A. Schlattl, A. M. Stutz, V. Benes, and J. O. Korbel, "DELLY: structural variant discovery by integrated paired-end and split-read analysis," *Bioinformatics,* vol. 28, pp. i333-i339, Sep 15 2012.

[8] J. Wang, C. G. Mullighan, J. Easton, S. Roberts, S. L. Heatley, J. Ma, *et al.*, "CREST maps somatic structural variation in cancer genomes with base-pair resolution," *Nat Methods,* vol. 8, pp. 652-4, Aug 2011.

[9] A. Christoforides, J. D. Carpten, G. J. Weiss, M. J. Demeure, D. D. Von Hoff, and D. W. Craig, "Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs," *BMC Genomics,* vol. 14, p. 302, 2013.

[10] A. R. Quinlan, R. A. Clark, S. Sokolova, M. L. Leibowitz, Y. Zhang, M. E. Hurles, *et al.*, "Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome," *Genome Res,* vol. 20, pp. 623-35, May 2010.

[11] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, *et al.*, "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Res,* vol. 20, pp. 1297-303, Sep 2010.

[12] C. Bartenhagen and M. Dugas, "RSVSim: an R/Bioconductor package for the simulation of structural variations," *Bioinformatics,* vol. 29, pp. 1679-81, Jul 1 2013.

[13] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, *et al.*, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics,* vol. 25, pp. 2078-9, Aug 15 2009.

[14] H. Li and R. Durbin, "Fast and accurate long-read alignment with Burrows-Wheeler transform," *Bioinformatics,* vol. 26, pp. 589-95, Mar 1 2010.

[15] The Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumours," *Nature,* vol. 490, pp. 61-70, Oct 4 2012.