Copy number variation detection using partial alignment information

Fatima Zare*, Sardar Ansari†§, Kayvan Najarian†‡§¶ and Sheida Nabavi*

*Department of Computer Science and Engineering, University of Connecticut, Storrs, Connecticut, 06269
‡Department of Emergency Medicine, §Department of Computational Medicine and Bioinformatics,
¶Department of Electrical Engineering and Computer Science,
† University of Michigan, Ann Arbor, Michigan, 48109
Email: *{fatemeh.zare,sheida.nabavi}@uconn.edu, †{sardara,kayvan}@umich.edu

***Abstract*—There are several approaches for copy number variation (CNV) detection using next-generation sequencing data. Among them, read-depth based methods have become widely used, especially for targeted and whole exome sequencing data. However, read-depth based approaches suffer from noise and biases and also poor breakpoint detection. In this work, we present a novel efficient segmentation algorithm that integrates information from partially mapped (soft-clipped and split) reads with read depth data for more precise CNV detection. The proposed method employs an efficient implementation of the solution to the change-point optimization problem, Taut String, to smooth the read depth data and to generate piecewise constant signals as CNV segments. Using simulated and real data, we show that our proposed method has a much faster runtime and can improve the sensitivity of CNV detection compared to the commonly used CBS method.**

Keywords*—*Copy number variation, soft-clipped read, split read, Breakpoints

I. Introduction

Copy number variations (CNVs) are an important type of genomic structural variations (SVs) and have been shown to associate with a variety of human diseases such as schizophrenia, autism, intellectual disability and cancer [1, 2]. CNVs are defined as deletions and multiplication of segments of a genome and vary in size between 50 bp and one whole chromosome arm [3, 4]. Whole exome sequencing (WES) provides a rapid and cost-effective sequencing for identifying genomic aberrations including CNVs and recently has become an effective and efficient alternative to whole genome sequencing (WGS) [5–7] , especially for clinical applications. However, due to biases and noise which are produced during the exome capturing procedure, the CNV detection tools for WGS data are not appropriate for WES data. Besides, the limitations of current CNV detection tools is another motive for developing new robust methods [8, 9] for WES data. Several tools have been developed for identifying SVs using sequencing data. These tools can be categorized into the de novo assembly, split-reads, read-pair, and read-depth or combination of different approaches [10, 11]. The de novo assembly based methods assemble reads without a genomic reference sequence and then call SVs using the difference between the assembled sequence and the genome sequence [11–14]. Split-reads approaches use unmapped reads (which their mates are mapped) or soft-clipped reads (partially aligned reads) to detect breakpoints. These types of reads span the breakpoint of SVs event and can carry information about the precise position of SVs' breakpoints. The methods that use soft-clipped reads can detect breakpoints using the mapping location where soft-clipping occurs [10], [15–23]. Read-pair strategies use the distance and orientation of paired-end reads to detect several types of SV events [24–27]. For detecting CNVs using WES data, the paired-read approach is not applicable as exons are generally smaller than insert sizes [5]. The read-depth approach is the commonly used method for identifying CNVs. The main hypothesis behind read-depth is that the read coverage is correlated with the copy number. Read-depth based methods use changes in read depths to detect CNVs regions. Amplified regions show higher read depths while deleted regions are associated with lower read depths compared to normal regions [5, 6, 9, 11, 28]. To have more accurate CNV detection, a combination of different approaches is suggested that also help to detect CNVs' breakpoints with base-pair resolution. SRBreak [29] and ClipCrop [22] combine split-reads and read-depth approaches to detect breakpoints. These tools extract soft-clipped reads information from Cigar strings of Bam files. Also, SVMerge [30] combines read-pair and read-depth approaches to detect SVs and then uses an assembler to improve breakpoint detection accuracy. There are also several tools that combine read-depth, paired-read and split-reads information to detect breakpoint positions [16, 20, 31, 32]. Read-depth data can be represented as a sparse vector with a limited number of non-zero values for CNV segments and can be considered as a piece-wise constant (PWC) signal [33]. For detecting a PWC signal, an isotonic regression can be used to estimate the change-points of the signal [34–37]. Isotonic regression is a convex optimization problem and is used for non-parametric estimation of probability distributions. Taut String is an efficient implementation of total variation denoising that is proposed to solve isotonic regression in linear time. In this work, for improving the sensitivity of CNV detection using read-depth signals, first, we estimate the potential location of breakpoints using split and soft-clipped reads. Then, using the estimated locations as prior knowledge for solving the change-point optimization problem, we take advantage of Taut String implementation to detect CNV segments efficiently.

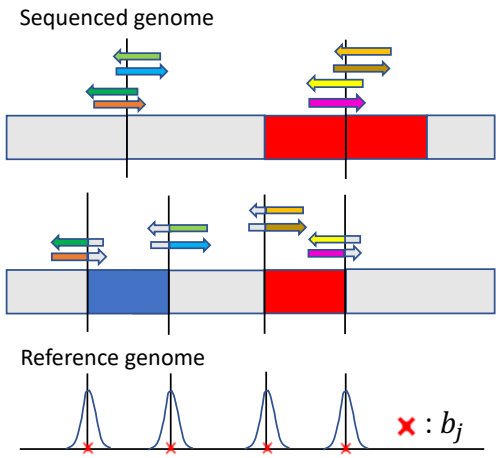

Fig. 1: Estimating potential CNVs breakpoints using soft-clipped reads. Red shows an amplified region and blue shows a deleted region. Gray parts of reads are not aligned to the reference genome

II. Method

The proposed method is divided into two parts, preprocessing and segmentation.

A. PreProcessing

1) Generating base-level read-depth signal : In this part, first, using BEDTools [38] we calculate base-level read depth for each of target exons for both sample and control data. Then, the ratio between read depths of the sample and control for each exon is multiplied by the ratio of the total sums of read depths of the sample and control in all the target exons to correct the imbalance library size effect between sample and control [39, 40] as in equations (1) and (2).

$$r_i = t_i/c_i \tag{1}$$

$$r'_i = r_i \frac{\sum_{i=1}^{e} c_i}{\sum_{i=1}^{e} t_i} \tag{2}$$

where t_i and c_i are the ith target exon's read depth for sample and control, respectively; and r'_i is the corrected ratio between them for ith target exon for $1 \leq i \leq e$, where e is the number of bases of the exon.

2) Estimating the CNVs breakpoints using partially aligned reads: BWA aligner [41] can output unmapped reads and soft-clipped alignments in the alignment file [42]. In this study, we use the partially mapped reads to estimate the potential locations of breakpoints. We use Cigar string to identify soft-clipped reads. Then, using BEDTools, we identify the potential breakpoints locations based on the partially mapped reads coverage. Figure 1 shows how soft-clipped reads can be used to detect CNVs breakpoints (b_js). We use potential breakpoints to adjust the segmentation method's (Taut String) tuning parameter, as described in the following section.

B. Segmentation

In this part, we apply Taut String to read-depth signals to smooth the signal and detect a PWC signal. Then using a sliding window technique, we fine tune the Taut String's output to detect CNVs segments.

1) Taut String to detect PWC signal: The identification of CNVs from the read-depth signal r'_i $(1 \leq i \leq e)$ can be considered as a change-point detection problem. Many methodologies are used to solve the change-point detection problem. Among them, total variation based regularization methods are an effective approach to remove noise and detect change points [33, 43]. In this work, for estimating CNVs segments, we use an efficient non-linear denoising method based on the total variation denoising for one-dimensional discrete signals [34, 35, 44, 45] that can preserve edges and narrow segments. The total variation denoising can identify local extreme values in very noisy data by estimating a piecewise constant signal. Given a noisy signal $\mathbf{r'} = (r'_1, r'_2, \ldots, r'_e)$, the goal is to estimate the denoised signal $\mathbf{f}$ which minimizes the following equation for some regularization parameter $\epsilon \geq 0$.

$$\min_f \frac{1}{2} \sum_{i=1}^{e} |r'_i - f_i|^2 + \epsilon \sum_{i=1}^{e-1} |f_{i+1} - f_i| \tag{3}$$

where the first term measures the fitting error between the noisy signal r'_i and smoothed signal f_i, and the second term measures the penalty caused by the change-point between f_i and f_{i+1} using a sparsity-inducing regularizer (ϵ). In fact, f can be computed efficiently using the Taut String method, and we will write $f = TS(\mathbf{r}, \epsilon)$. The challenging part is selecting an appropriate value for ϵ. Taut String defines a vector of running sums $R'_i = \sum_{u=1}^{i} r'_u$, $1 \leq i \leq e$. If $\mathbf{F} = TS(\mathbf{R'}, \epsilon)$ then $\mathbf{f} = D(\mathbf{F})$ can approximate r' as well, where D is the differentiation operation, because $\|\mathbf{r'} - \mathbf{f}\|_2$ is minimized under the restriction $\|\mathbf{R'} - \mathbf{F}\|_\infty \leq \epsilon$ [8, 46, 47].

Therefore, for a fixed $\epsilon \geq 0$, Taut String tries to find $\mathbf{F}$, such that:

$$\|\mathbf{R'} - \mathbf{F}\|_\infty = \max_i \{|R'_i - F_i|\} \leq \epsilon \tag{4}$$

To satisfy equation (4), $\mathbf{F}$ should minimize the norm 2 of a differentiation operation D and norm 1 of the second derivative of $\mathbf{F}$ as the optimization objective:

$$\|D(\mathbf{F})\|_2 = \sqrt{\sum_{i=1}^{e-1} (F_{i+1} - F_i)^2}, \tag{5}$$

$$\|D^* D(\mathbf{F})\|_1 = |F_2 - F_1| + \sum_{i=2}^{e-1} |F_{i-1} - 2F_i + F_{i+1}| + |F_n - F_{n-1}| \tag{6}$$

where $D^* : R^{n-1} \to R^n$ is dual to $D : R^n \to R^{n-1}$:

$$D(F) = (F_2 - F_1, F_3 - F_2, \ldots, F_n - F_{n-1}) \tag{7}$$

$$D^*(d_1, d_2, \ldots, d_{n-1}) = (-d_1, d_1 - d_2, \ldots, d_{n-2} - d_{n-1}, d_{n-1}). \tag{8}$$

In the end, we apply the differentiation operation to the output of Taut String to obtain **f**.

2) Optimal tuning for the regularization parameter ϵ: Finding an appropriate value for the regularization parameter ϵ is the challenging part of our algorithm. For having a more accurate estimation of the PWC signal, we need to adjust the requlalrization parameter adaptively. Using soft-clipped reads, we estimate the potential location of CNV regions and then try to find an optimal regularization parameter for each potential region. To determine the regularization parameter ϵ we applied the Schwartz information criterion (SIC) on each estimated region [33, 48]. We calculate SIC for k number of ϵ for each estimated region. The SIC at ϵ_k is determined as:

$$SIC(\epsilon_k) = np \times \ln(m) + \frac{\sum_{l=1}^{m} (r'_l - f_l)^2}{\sigma^2} \tag{9}$$

where np is the number of identified pieces by Taut String in each estimated region , m is the length of the signal, σ^2 is the variance of noise which can be determined as the median of the standard deviation of all detected pieces by Taut String. The optimal ϵ is obtained as:

$$\hat{\epsilon} = \arg\min_{k} SIC(\epsilon_k). \tag{10}$$

Using the optimal ϵ for each region, we smooth the read depth data to a PWC signal.

3) Detecting CNVs segments using the sliding window approach: Taut String provides a PWC signal, but the signal contains many adjacent constant pieces that their amplitudes are very close. We merge these pieces to identify CNV segments using a sliding window approach. The sliding window approach with linear time complexity $\mathcal{O}(n)$ can transform the signal to a piecewise linear approximation and mainly used in time series segmentation methods for detecting change points [49, 50]. The algorithm tries to grow a segment until it exceeds an error bound defined by users. The point that the error bound exceeds its threshold is considered as CNV's breakpoints. At the end, we calculate the medians of the log_2 ratios of pieces between each pair of breakpoints as copy number values of CNV segments.

III. DATASETS

To investigate the performance of the proposed method, we used three sets of datasets: 1) simulated readcount data, 2) simulated sequencing data, and 3) real data.

A. Simulated read count datasets

We used the detected CNVs segments from chromosome 2 of real data, obtained by applying Varscan2 [51] and CBS segmentation [52] as the known CNV segments. After sampling from the known CNV segments of 10 real datasets at 100 bp genomic distances, we added noise to simulate noisy read-depth signals.

B. Simulated sequencing datasets

We have also used a CNV simulator, called CNV-Sim (https://github.com/NabaviLab/CNV-Sim) to simulate WES data with known CNVs to evaluate the performance of the proposed method. We generated ten paired-end WES datasets with the read length of 100 bp for chromosome 1. We used the BWA tool [41] to align short reads to the reference genome (hg19) and generated BAM files. Then, using BEDTools [38], we calculated the per-base depth of coverage to generate read-depth data for these simulated sequencing data.

C. Real datasets

In this study, we used five breast cancer tumor and matched normal WES data. The aligned BAM files of these five tumor-normal pairs were downloaded from the Cancer Genomics Hub (CGHub), https://cghub.ucsc.edu/index.html. We also used array-based CNV data from the same five tumor samples as a benchmark for the CNV detection tools evaluation, downloaded from the TCGA data portal (https://portal.gdc.cancer.gov/projects/TCGA-BRCA). The BEDTools suit was utilized to generate per-base read-depth signals from the aligned sequencing data.

IV. RESULTS AND DISCUSSION

A. Results on simulated datasets

In this section, we evaluated the effect of the employing estimated locations of breakpoints in detecting CNVs. We applied our proposed method to ten simulated signals with a signal to noise ratio of 7. We compared the sensitivity and false discovery rate (FDR) of detecting CNVs for the case that we know the exact locations of breakpoints with cases that we have approximate locations of breakpoints as it happens in real data. We considered several approximations to the exact breakpoints ranging from 100 bp to 500 bp. We considered 0.1 as the sliding window error bound. We applied a segment-based comparison [9] to call true positive (TP), false negative (FN) and false positive (FP). We used the "GenomicRanges" R package from Bioconductor [53] to calculate overlapping regions between detected CNVs and benchmark CNVs. The threshold of ± 0.1 for log2ratio and 50% overlap were used for calling CNV segments. Figure 2 shows the Sensitivity and FDR using a segment-based approach for different distances from the exact locations of breakpoints. It is evident that the exact locations of breakpoints can provide the best sensitivity and FDR; and closer the approximations to the exact locations, the better the performance is. However, estimations with less than 300 bp difference from the exact locations can provide sensitivities higher than 0.95 and FDRs less than 0.05. This results indicate that the proposed method can perform well for simulated data when we have close approximations of breakpoints.

Also, we applied the circular binary segmentation (CBS) [52] method, which is a well-established segmentation method, to the simulated datasets to compare its performance with that of our proposed method. CBS showed sensitivities from 0.84% to 0.93% and FDRs from 0.15% to 0.09% for different values of undo.SD parameter ranging from 12 to 4. It is clear that the proposed method outperforms CBS in detecting CNVs segments.

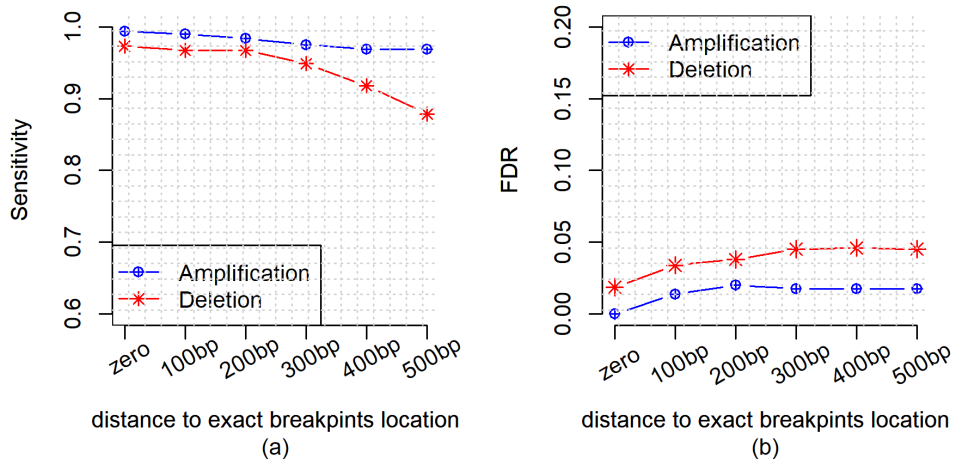

Fig. 2: The performance of the proposed method in detecting CNVs considering the prior knowledge about CNVs breakpoints in terms of a) sensitivity b) FDR, with error bound equal to 0.1 and CNV $thr = 0.2$

TABLE I: Overall Performance of CNV detection Methods Using the simulated WES data generated by CNV-Sim Data

Method	**Amplification**			**Deletion**		
	Sensitivity	**FDR**	**Specificity**	**Sensitivity**	**FDR**	**Specificity**
Our proposed method	95.76%	27.24%	99.73%	88.74%	35.05%	**99.67**%
VarScan2	**82.98**%	11.88%	99.91%	80.55%	18.73%	99.88%

TABLE II: Overall Performance of the Proposed CNV detection Method Using the Real WES data

Methods	**Amplification**			**Deletion**		
	Sensitivity	**FDR**	**Specificity**	**Sensitivity**	**FDR**	**Specificity**
Our proposed method	80.08%	18.20%	83.31%	76.85%	18.56%	86.17%
VarScan2	65.67%	15.49%	88.70%	76.56%	11.43%	88.44%
ExomeCNV	84.02%	11.83%	90.85%	90.43%	14.95%	86.85%

B. Results on simulated sequencing data

We compared the performance of our proposed method with that of VarScan2 with CBS, using simulated sequencing data. We used the gene-based approach to calculate sensitivities, FDRs, and specificities [9]. We used the "cghMCR" R package from Bioconductor [54] to identify CNV genes using Refseq gene identifications. Thresholds of ± 0.1 were used to call CNV genes. The average of sensitivities, FDRs, and specificities for detecting amplified and deleted genes are shown in Table I. We observe that tools show higher FDRs in detecting deleted genes compared to detecting amplified genes. The proposed method shows higher sensitivity compared to VarScan2 where its FDR is poorer, but it has comparable specificity compared to VarScan2. Higher sensitivities indicate the ability of our proposed method to detect more CNV regions including short and focal CNVs. Need to mention that the simulator does not provide unmapped paired-end reads. Therefore, only soft-clipped reads were used to compute the coverage of the split reads and finding the approximate of breakpoints' locations. Using only soft-clipped read provides partial information about breakpoints. It can be the reason for the high FDRs of the proposed method. Having more information about the location of breakpoints will decrease False detection and help to improve the performance of the segmentation method as shown by using simulated read-depth data.

Detection of short segments is a challenging problem for current CNV detection tools. We analyzed the FN, FP and TP CNV segments obtained by our method and VarScan2, regarding their lengths. We observed that the power of our proposed method for detecting short CNVs is higher compared to VarScan2. The proposed method that employs base pair level read depth data and utilizes an edge protection change-point detection method can detect very short CNV segments.

C. Results on real datasets

VarScan2 and ExomeCNV are two commonly used CNV detection tools with strong performances using WES data [9]. They are able to detect somatic aberration and use the read-depth approach. Both of them use CBS segmentation in their segmentation part. The average sensitivity, specificity, and FDR of these tools on real breast cancer WES data are

shown in Table II. Thresholds of ± 0.1 were used to call CNV genes. Table II shows that the proposed method has better or comparable sensitivities compared to the other two methods. However, its FDRs are poorer. There are two main reasons for this poor FDR. First, some of the detected FPs are small CNV regions which VarScan2 and ExomeCNV are not able to detect them, and they are not included in the benchmarks as the benchmarks are from array-based technology that contains larger CNVs. The second reason is that our information about the approximate location of breakpoints is not complete. The BAM files provided by TCGA do not include the unmapped paired-end read. The information about the breakpoint was extracted only based on the soft-clipped read. Also, the soft-clipped read information is very noisy because partial alignment can be due to other reasons such as sequencing error or indels besides CNVs. Defiantly including the coverage of split reads from unmapped read will boost the performance of the proposed method. Here we showed that even with incomplete information the proposed method performs well with high sensitivity and specificity and moderate FDR.

D. Runtime Comparision

We compared the runtime of the proposed method with that of the CBS. Many CNV detection tools use CBS for calling CNV segments. CBS uses an iterative algorithm based on the variance of the data.

One of the main drawbacks of the CBS for WES or WGS data is its running time. Using simulated WES datasets for Chr1 and running on a 64-bit Windows 10 Operating System, with Intel Core i7-7500U 2.7 GHz CPU and 16 GB DDR4 memory, CBS takes in average 600 seconds to 900 seconds for different values of undo.SD parameter (ranging from 12 to 4). However, it takes in average about 20 seconds for the proposed method. The main reason for the lower overall runtime of our algorithm is due to employing the Taut String algorithm in our pipeline. The complexity of Taut String algorithm is linear in time. It is a non-iterative in-place method for finding the solution to an optimization problem. Finding the optimal regularization parameter for a large dataset can be computationally expensive. However, because we divided the entire read depth data to regions based on the protentional breakpoint, finding the optimal parameter for each region using SIC is very fast.

V. CONCLUSION

In this study, we developed a new segmentation method to identify CNVs that is based on the combination of read-depth and split-read approaches. Our proposed method can simultaneously remove noise, preserve edges, and generate piecewise constant signals. From a signal processing point of view, the read-depth signal is sparse, discrete, and piecewise constant. Therefore, using an edge protection change-point detection approach such as Taut String can result in a better CNV detection. However, the performance of Taut String depends on its regularization parameter that is very challenging to find its optimal value. In this work, we used the prior knowledge about the potential locations of breakpoints, extracted from the partially aligned reads (soft-clipped reads) to obtain the optimal adaptive values for the regularization parameter. The proposed method has several advantages and limitations. It is an efficient segmentation method. It can detect very short and focal CNV segments. And, it provides very high sensitivities in detecting CNVs. However, its FDR is not very low, and it requires complete information about potential breakpoints. For our future works, we will work on realigning unmapped and soft-clipped reads to have more robust information about the potential locations of breakpoints and optimal selection of the regularization parameter to improve the performance of our segmentation method.

VI. ACKNOWLEDGEMENT

This study was supported by a grant from the National Institutes of Health (NIH, R00LM011595, PI: Nabavi).

REFERENCES

[1] A. Shlien and D. Malkin, "Copy number variations and cancer," *Genome medicine*, vol. 1, no. 6, p. 62, 2009.

[2] N. Zhang, M. Wang, P. Zhang, and T. Huang, "Classification of cancers based on copy number variation landscapes," *Biochimica et Biophysica Acta (BBA) - General Subjects*, vol. 1860, no. 11, pp. 2750–2755, Nov. 2016. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0304416516302082

[3] B. Nowakowska, "Clinical interpretation of copy number variants in the human genome," *J. Appl. Genet.*, vol. 58, no. 4, pp. 449–457, Nov. 2017.

[4] H. Zhang, A. F. B. Zeidler, W. Song, C. M. Puccia, E. Malc, P. W. Greenwell, P. A. Mieczkowski, T. D. Petes, and J. L. Argueso, "Gene copy-number variation in haploid and diploid strains of the yeast Saccharomyces cerevisiae," *Genetics*, vol. 193, no. 3, pp. 785–801, Mar. 2013.

[5] J. F. Sathirapongsasuti, H. Lee, B. A. J. Horst, G. Brunner, A. J. Cochran, S. Binder, J. Quackenbush, and S. F. Nelson, "Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV," *Bioinformatics*, vol. 27, no. 19, pp. 2648–2654, Oct. 2011.

[6] R. Yao, C. Zhang, T. Yu, N. Li, X. Hu, X. Wang, J. Wang, and Y. Shen, "Evaluation of three read-depth based CNV detection tools using whole-exome sequencing data," *Mol Cytogenet*, vol. 10, p. 30, 2017.

[7] G. H. Reference, "What are whole exome sequencing and whole genome sequencing?" [Online]. Available: https://ghr.nlm.nih.gov/primer/testing/sequencing

[8] F. Zare, S. Ansari, K. Najarian, and S. Nabavi, "Noise cancellation for robust copy number variation detection using next generation sequencing data." IEEE, Nov. 2017, pp. 230–236. [Online]. Available: http://ieeexplore.ieee.org/document/8217654/

[9] F. Zare, M. Dow, N. Monteleone, A. Hosny, and S. Nabavi, "An evaluation of copy number variation detection tools for cancer using whole exome sequencing data," *BMC bioinformatics*, vol. 18, no. 1, p. 286, 2017.

[10] Y. Wu, L. Tian, M. Pirastu, D. Stambolian, and H. Li, "MATCHCLIP: locate precise breakpoints for copy number variation using CIGAR string by matching soft clipped reads," *Frontiers in Genetics*, vol. 4, 2013. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fgene.2013.00157/abstract

[11] Y. Feng, D. Chen, and L.-J. C. Wong, "Detection of Copy Number Variations (CNVs) Based on the Coverage Depth from the Next Generation Sequencing Data," in *Next Generation Sequencing Based Clinical Molecular Diagnosis of Human Genetic Disorders*, L.-J. C. Wong, Ed. Cham: Springer International Publishing, 2017, pp. 13–22. [Online]. Available: http://link.springer.com/10.1007/978-3-319-56418-0_2

[12] E. B. Daniel R. Zerbino, "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs."

[13] R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, K. Kristiansen, S. Li, H. Yang, J. Wang, and J. Wang, "De novo assembly of human genomes with massively parallel short read sequencing," *Genome Research*, vol. 20, no. 2, pp. 265–272, Feb. 2010. [Online]. Available: http://genome.cshlp.org/cgi/doi/10.1101/gr.097261.109

[14] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones, and I. Birol, "ABySS: A parallel assembler for short read sequence data," *Genome Research*, vol. 19, no. 6, pp. 1117–1123, Jun. 2009. [Online]. Available: http://genome.cshlp.org/cgi/doi/10.1101/gr.089532.108

[15] 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, no. 21, pp. 2865–2871, Nov. 2009. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp394

[16] T. Rausch, T. Zichner, A. Schlattl, A. M. Stütz, V. Benes, and J. O. Korbel, "DELLY: structural variant discovery by integrated paired-end and split-read analysis," *Bioinformatics*, vol. 28, no. 18, pp. i333–i339, Sep. 2012.

[17] J. Zhang and Y. Wu, "SVseq: an approach for detecting exact breakpoints of deletions with low-coverage sequence data," *Bioinformatics*, vol. 27, no. 23, pp. 3228–3234, Dec. 2011. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr563

[18] Y. Jiang, Y. Wang, and M. Brudno, "PRISM: Pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants," *Bioinformatics*, vol. 28, no. 20, pp. 2576–2583, Oct. 2012. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts484

[19] A. Abyzov and M. Gerstein, "AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision," *Bioinformatics*, vol. 27, no. 5, pp. 595–603, Mar. 2011. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq713

[20] S. N. Hart, V. Sarangi, R. Moore, S. Baheti, J. D. Bhavsar, F. J. Couch, and J.-P. A. Kocher, "SoftSearch: Integration of Multiple Sequence Features to Identify Breakpoints of Structural Variations," *PLoS ONE*, vol. 8, no. 12, p. e83356, Dec. 2013. [Online]. Available: http://dx.plos.org/10.1371/journal.pone.0083356

[21] Z. Zhang, J. Wang, J. Luo, X. Ding, J. Zhong, J. Wang, F.-X. Wu, and Y. Pan, "Sprites: detection of deletions from sequencing data by re-aligning split reads," *Bioinformatics*, vol. 32, no. 12, pp. 1788–1796, Jun. 2016. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw053

[22] S. Suzuki, T. Yasuda, Y. Shiraishi, S. Miyano, and M. Nagasaki, "ClipCrop: a tool for detecting structural variations with single-base resolution using soft-clipping information," *BMC Bioinformatics*, vol. 12, no. Suppl 14, p. S7, 2011. [Online]. Available: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-S14-S7

[23] J. Schröder, A. Hsu, S. E. Boyle, G. Macintyre, M. Cmero, R. W. Tothill, R. W. Johnstone, M. Shackleton, and A. T. Papenfuss, "Socrates: identification of genomic rearrangements in tumour genomes by re-aligning soft clipped reads," *Bioinformatics*, vol. 30, no. 8, pp. 1064–1072, Apr. 2014. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt767

[24] L. Tattini, R. D'Aurizio, and A. Magi, "Detection of Genomic Structural Variants from Next-Generation Sequencing Data," *Front Bioeng Biotechnol*, vol. 3, p. 92, 2015.

[25] K. Chen, J. W. Wallis, M. D. McLellan, D. E. Larson, J. M. Kalicki, C. S. Pohl, S. D. McGrath, M. C. Wendl, Q. Zhang, D. P. Locke, X. Shi, R. S. Fulton, T. J. Ley, R. K. Wilson, L. Ding, and E. R. Mardis, "BreakDancer: an algorithm for high-resolution mapping of genomic structural variation," *Nat. Methods*, vol. 6, no. 9, pp. 677–681, Sep. 2009.

[26] F. Hormozdiari, C. Alkan, E. E. Eichler, and S. C. Sahinalp, "Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes," *Genome Res.*, vol. 19, no. 7, pp. 1270–1278, Jul. 2009.

[27] J. O. Korbel, A. Abyzov, X. J. Mu, N. Carriero, P. Cayting, Z. Zhang, M. Snyder, and M. B. Gerstein, "PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data," *Genome Biol.*, vol. 10, no. 2, p. R23, Feb. 2009.

[28] A. Abyzov, A. E. Urban, M. Snyder, and M. Gerstein, "Cnvnator: An approach to discover, genotype and characterize typical and atypical cnvs from family and population genome sequencing," *Genome research*, pp. gr–114 876, 2011.

[29] H. T. Nguyen, J. Boocock, T. R. Merriman, and M. A. Black, "SRBreak: A Read-Depth and Split-Read Framework to Identify Breakpoints of Different Events Inside Simple Copy-Number Variable Regions," *Frontiers in Genetics*, vol. 7, Sep. 2016. [Online]. Available: http://journal.frontiersin.org/Article/10.3389/fgene.2016.00160/abstract

[30] K. Wong, T. M. Keane, J. Stalker, and D. J. Adams, "Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly," *Genome Biology*, vol. 11, no. 12, p. R128, 2010. [Online]. Available: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-12-r128

[31] R. M. Layer, C. Chiang, A. R. Quinlan, and I. M. Hall, "LUMPY: a probabilistic framework for structural variant discovery," *Genome Biology*, vol. 15, no. 6, p. R84, 2014. [Online]. Available: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-6-r84

[32] J. Kim and K. Reinert, "Vaquita: Fast and Accurate Identification of Structural Variation Using Combined Evidence," 2017.

[33] J. Duan, J.-G. Zhang, H.-W. Deng, and Y.-P. Wang, "CNV-TV: A robust method to discover copy number variation from short sequencing reads," *BMC Bioinformatics*, vol. 14, no. 1, p. 150, 2013. [Online]. Available: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-150

[34] B. Meeuwissen, "Forecasting with Taut Strings," p. 36.

[35] L. Dümbgen, A. Kovac, and others, "Extensions of smoothing via taut strings," *Electronic Journal of Statistics*, vol. 3, pp. 41–75, 2009.

[36] H. Cho and P. Fryzlewicz, "Multiscale interpretation of taut string estimation and its connection to Unbalanced Haar wavelets," *Statistics and Computing*, vol. 21, no. 4, pp. 671–681, Oct. 2011, arXiv: 1611.08634. [Online]. Available: http://arxiv.org/abs/1611.08634

[37] N. C. Overgaard, "On the Taut String Interpretation of the One-dimensional Rudin-Osher-Fatemi Model: A New Proof, a Fundamental Estimate and Some Applications," *arXiv:1710.10985 [cs, eess]*, Oct. 2017, arXiv: 1710.10985. [Online]. Available: http://arxiv.org/abs/1710.10985

[38] A. R. Quinlan and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, no. 6, pp. 841–842, 2010.

[39] J. Kong, J. Shin, J. Won, K. Lee, U. Lee, and J. Yoon, "ExCNVSS: A Noise-Robust Method for Copy Number Variation Detection in Whole Exome Sequencing Data," *BioMed Research International*, vol. 2017, pp. 1–11, 2017. [Online]. Available: https://www.hindawi.com/journals/bmri/2017/9631282/

[40] J. Li, R. Lupat, K. C. Amarasinghe, E. R. Thompson, M. A. Doyle, G. L. Ryland, R. W. Tothill, S. K. Halgamuge, I. G. Campbell, and K. L. Gorringe, "Contra: copy number analysis for targeted resequencing," *Bioinformatics*, vol. 28, no. 10, pp. 1307–1313, 2012.

[41] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, Jul. 2009. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp324

[42] D. M. Bickhart, J. L. Hutchison, L. Xu, R. D. Schnabel, J. F. Taylor, J. M. Reecy, S. Schroeder, C. P. Van Tassell, T. S. Sonstegard, and G. E. Liu, "RAPTR-SV: a hybrid method for the detection of structural variants," *Bioinformatics*, vol. 31, no. 13, pp. 2084–2090, Jul. 2015. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv086

[43] H. Derksen, "A general theory of singular values with applications to signal denoising," *arXiv:1705.10881 [cs, math, stat]*, May 2017, arXiv: 1705.10881. [Online]. Available: http://arxiv.org/abs/1705.10881

[44] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1-4, pp. 259–268, Nov. 1992. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/016727899290242F

[45] Junbo Duan, Ji-Gang Zhang, J. Lefante, Hong-Wen Deng, and Yu-Ping Wang, "Detection of copy number variation from next generation sequencing data with total variation penalized least square optimization." IEEE, Nov. 2011, pp. 3–12. [Online]. Available: http://ieeexplore.ieee.org/document/6112348/

[46] A. Belle, S. Ansari, M. Spadafore, V. A. Convertino, K. R. Ward, H. Derksen, and K. Najarian, "A Signal Processing Approach for Detection of Hemodynamic Instability before Decompensation," *PloS one*, vol. 11, no. 2, p. e0148544, 2016.

[47] K. Najarian, A. Belle, K. Ward, and H. Derksen, "Early detection of hemodynamic decompensation using taut-string transformation," US Patent US9 974 488B2, May, 2018. [Online]. Available: https://patents.google.com/patent/US9974488B2/en

[48] G. Schwarz, "Estimating the Dimension of a Model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, Mar. 1978. [Online]. Available: https://projecteuclid.org/euclid.aos/1176344136

[49] E. Keogh, S. Chu, D. Hart, and M. Pazzani, "SEGMENTING TIME SERIES: A SURVEY AND NOVEL APPROACH," in *Series in Machine Perception and Artificial Intelligence*. WORLD SCIENTIFIC, Jun. 2004, vol. 57, pp. 1–21. [Online]. Available: http://www.worldscientific.com/doi/abs/10.1142/9789812565402_0001

[50] S. Aminikhanghahi and D. J. Cook, "A survey of methods for time series change point detection," *Knowledge and information systems*, vol. 51, no. 2, pp. 339–367, 2017.

[51] D. C. Koboldt, Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding, and R. K. Wilson, "VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing," *Genome research*, vol. 22, no. 3, pp. 568–576, 2012.

[52] A. B. Olshen, E. Venkatraman, R. Lucito, and M. Wigler, "Circular binary segmentation for the analysis of array-based DNA copy number data," *Biostatistics*, vol. 5, no. 4, pp. 557–572, 2004.

[53] "GenomicRanges." [Online]. Available: http://bioconductor.org/packages/GenomicRanges/

[54] "cghMCR." [Online]. Available: http://bioconductor.org/packages/cghMCR/