Published in final edited form as:

ACM BCB. 2019 September ; 2019: 423–428. doi:10.1145/3307339.3342181.

Copy Number Variation Detection Using Total Variation

Fatima Zare, Sheida Nabavi*

University of Connecticut Storrs, Connecticut 06269

Abstract

Next-generation sequencing (NGS) technologies offer new opportunities for precise and accurate identification of genomic aberrations, including copy number variations (CNVs). For high-throughput NGS data, using depth of coverage has become a major approach to identify CNVs, especially for whole exome sequencing (WES) data. Due to the high level of noise and biases of read-count data and complexity of the WES data, existing CNV detection tools identify many false CNV segments. Besides, NGS generates a huge amount of data, requiring to use effective and efficient methods. In this work, we propose a novel segmentation algorithm based on the total variation approach to detect CNVs more precisely and efficiently using WES data. The proposed method also filters out outlier read-counts and identifies significant change points to reduce false positives. We used real and simulated data to evaluate the performance of the proposed method and compare its performance with those of other commonly used CNV detection methods. Using simulated and real data, we show that the proposed method outperforms the existing CNV detection methods in terms of accuracy and false discovery rate and has a faster runtime compared to the circular binary segmentation method.

Keywords

Copy Number Variation; Next Generation Sequencing; Whole Exome Sequencing; Signal Processing; Total Variation; Taut String

1 INTRODUCTION

An important application of next generation sequencing (NGS) is detection of copy number variations (CNVs) [12, 27, 28]. CNVs are an important type of structural variations that result in either gain (amplification) or loss (deletion) of genomic regions. To identify CNVs, whole exome sequencing (WES) and whole genome sequencing (WGS), have become primary strategies for NGS. CNV detection tools developed for WGS data are not appropriate for WES data. This is because, in WES, sequencing data are available only for exonic regions, and exome capture procedures introduce more biases and noise. Therefore, it

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

*Corresponding author. This study was supported by a grant from the National Institutes of Health (NIH, R00LM011595, PI: Nabavi). sheida.nabavi@uconn.edu.

is necessary to build a robust and precise model to detect CNVs for WES data. Moreover, the key feature of NGS is that it generates huge amounts of data (usually at the scale of gigabytes), which requires to use efficient methods. The depth of coverage (DOC) approach is the most appropriate method to identify CNVs for WES data. The main hypothesis behind the DOC method is that the read-count is correlated with the copy number at a genomic region. Amplified regions show higher read-counts, while deleted regions have lower read-counts compared to normal regions [35].

In general, the DOC-based tools for CNVs detection are divided into two major steps: 1) preprocessing, and 2) segmentation. In the preprocessing step, noise and biases are reduced from a read-count signal, and in the segmentation part, CNV segments are identified by merging the regions with similar read-count values. Circular binary segmentation (CBS) [19] and hidden Markov model (HMM) [7] are the most widely used segmentation algorithms in the existing CNV detection tools [37]. CBS locates the breakpoints recursively until the chromosomes are divided into segments with equal copy numbers that differ significantly from the copy numbers of their neighbor genomic regions. CBS is an effective segmentation method; however, it is slow when read-count data are very noisy. In HMM, the read-count windows are binned sequentially along the chromosome depending on defined states of amplification, deletion, and no CNVs. Then, CNV segments are identified by combining consecutive windows with the same states. Although the HMM algorithm is an efficient algorithm, it may not work well if the initial assumptions are not correct. In the DOC-based approach, CNV segments can be represented as piece-wise constant (PWC) signals [6]. Detecting PWC signals can be formulated as detecting change points. The total variation (TV) approach is a sparse-regularized optimization that have shown outstanding performances in estimating change points of a PWC signal [4, 5, 8]. Taut String is an efficient implementation of TV that is proposed to solve isotonic regression in linear time [8]. However, for highly noisy data, Taut String suffers from detecting multiple consecutive change points in the margin of a real change point that is called the staircase effect. This staircase effect results in false positives and limits the use of Taut String algorithm for segmentation [35].

In this work, we developed a novel efficient segmentation algorithm by modifying the Taut String algorithm to reduce the staircase effect for identifying more precise CNVs [1]. We have also developed a statistical approach based on the Pettitt test to assign significant values to detected change points for filtering out the low confidence CNV segments to further reduces false positive CNV segments. The codes are available at <https://github.com/NabaviLab/TSCNV>.

2 METHOD

2.1 Identifying CNV segments

Using read-count data, CNV segments can be considered as PWC signals where detecting CNV segments can be achieved by detecting change points in read-count data. A change point shows a sudden change in the statistics of a sequence of data points. In this section, first, we introduce a new efficient change points detection algorithm based on Taut String for accurate detection of CNV change points (breakpoints) using read-count data. Then, we

explain a very effective and efficient statistical test based on the Pettitt test to assign significant values to the detected CNV change points to remove false positives. Finally, using the detected high confidence change points, we call CNV segments from the read-count data. The overall block diagram of the proposed method is shown in Figure 1.

2.1.1 Iterative Taut String Algorithm for detecting change points.—The TV approach for one-dimensional (1D) discrete signals (TV-1D) computes the least square estimates of the desired data for some regularization parameter $\epsilon > 0$:

$$\min_{\mathbf{f} \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{r} - \mathbf{f}\|_2^2 + \epsilon \|\mathbf{D}\mathbf{f}\|_1, \quad (1)$$

where $\mathbf{r}$ is the observed data, $\mathbf{f}$ is the desired data, and $\mathbf{D}$ is the differencing matrix (all zeros except $D_{ii} = -1$ and $D_{i,i+1} = 1$ ($i = 1, \dots, n-1$)) [1]. Taut String is an efficient implementation of TV-1D that solve isotonic regression in linear time and provides a fast solution for minimizing the cost function in Equation (1). Introducing a dual variable $\mathbf{u} \in \mathbb{R}^n$, the authors in [1] show that Equation (1) has a dual form as:

$$\min_{\mathbf{u} \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{D}^T \mathbf{u}\|_2^2 - \mathbf{u}^T \mathbf{D}\mathbf{r}, \text{ s.t. } \|\mathbf{u}\|_\infty \leq \epsilon, \quad (2)$$

where $\mathbf{u}$ is the difference between the observed and desired signal for the optimization problem in Equation (1). Now, by changing and defining a new set of variables, we have:

$$\min_{\mathbf{F}} \sum_{i=1}^n \sqrt{1 + (F_{i-1} - F_i)^2}, \text{ s.t. } \|\mathbf{F} - \mathbf{R}\|_\infty \leq \epsilon, \quad (3)$$

where $R_i = \sum_{k=1}^i r_k$ and $F_i = \sum_{j=1}^i f_k$ are the cumulative sum of signal $\mathbf{r}$ and $\mathbf{f}$ values respectively and we have $F_0 = 0$, $R_n = F_n$. Once $\mathbf{F}$ is found, we can recover the solution for the original TV-1D problem in Equation (1) by observing that:

$$F_i - F_{i-1} = R_i - u_i - (R_{i-1} - u_{i-1}) = r_i - u_i + u_{i-1} = f_i. \quad (4)$$

According to $F_i = \sum_{j=1}^i f_k$, we can conclude that f_i is the slope of $\mathbf{F}$ between point i and $i+1$. So, f_i represents the derivative of $\mathbf{F}$ between point i and $i+1$. Similarly, we can interpret $\mathbf{r}$ and $\mathbf{R}$ in the same way as $\mathbf{f}$ and $\mathbf{F}$ [8]. The recovered signal $\mathbf{F}$ can be thought of a string between $R + \epsilon$ and $R - \epsilon$ that is pulled tight. Using the conventional Taut String algorithm for segmentation has two main limitations: the staircase effect and adjusting the regularization parameters. Rojas et al. argued that if the slopes of consecutive segments are increasing (or decreasing), TV-1D algorithm will detect two or more consecutive change points in the margin of a real change point, referred to the staircase effect since the signs of the consecutive changes are the same [20, 29, 30]. The staircase effect is worse when the level of noise is high (Figure 2(a) and 2(b)). Finding an appropriate value for the regularization parameter ϵ is another challenge for the Taut String algorithm. In fact, for having a more accurate estimation of a PWC signal, we need to adjust the regularization

parameter adaptively. A low value of ϵ causes detecting many small false segments and a high value of ϵ causes missing true segments as shown in Figure 2(c) and 2(d).

In this work, we propose a modified Taut String algorithm to address the staircase effect and choosing an appropriate regularization parameter problems as shown in Algorithm 1. It has been shown that the staircase event does not affect the first and last detected change points of TV-1D [29]. Therefore, our proposed algorithm, after detecting the first change point, makes Taut String to stop, stores the first change point, and resets again with a new start point. The algorithm repeats this procedure until it investigates the whole signal (Algorithm 1). Even though we are using the Taut String algorithm iteratively, the overall complexity is still linear in the time since at each iteration the algorithm stops at the first detected change point and start from this point. Figure 3 shows that the modified Taut String decreases the staircases. For a noisy signal with length n and standard deviation σ , the best ϵ can be chosen as $\sigma\sqrt{2\log n}$ [4, 5]. In our algorithm, we update ϵ based on the signal in each iteration. It means that after each pause, the algorithm applies Taut String on the rest of the signal with a new ϵ based on the new length and standard deviation of the signal, see Algorithm 1. Our developed algorithm reduces the staircase effect and at the same time weighs signal $\mathbf{R}$ with different regularization parameter in each iteration. So, after each pause, the algorithm calculates a new ϵ based on a shorter signal which promises a more accurate ϵ . In the end, the K potential change points $CP = (\tau_1, \dots, \tau_K)$ are derived by applying the differentiation operation to the detected signal $\mathbf{f}$.

2.1.2 Removing false positive change points.—The list of detected change points $CP = (\tau_1, \dots, \tau_K)$ contains true but also false change points. We developed a method to calculate the detected change points' p-values using the Pettitt test [18] to eliminate low confidence change points as false positives. The main advantages of this test are that it does not require any assumption on the distribution of data, and it provides a p-value to test the significance of a detected change point [18, 25]. The null hypothesis is no change in the distribution of a sequence of data, against the alternative that a change point exists. For a sequence of X with the length of T , the non-parametric statistic is defined as $L_T = \max |U_t - T|$, where $U_t, T = \sum_{i=1}^t \sum_{j=t+1}^T \text{sgn}(X_i - X_j)$. U_t, T is a statistic parameter used for analyzing the two sequences $X_1, \dots, X_t$ and $X_{t+1}, \dots, X_T$ arise from the same distribution. If the change point of the sequence X is located at L_T the significance probability of L_T is approximated with $p \cong 2 \times \exp\left(\frac{-6 \times L_T^2}{T^3 + T^2}\right)$.

The Pettitt test is used to detect a single change point of a signal. The developed method divides the detected PWC signal to pieces such that each piece contains a single change point. Then it applies the Pettitt test to each piece to compute the p-value of its change point. Users can keep the change points which have a p-value smaller than a user-defined critical level [3]. In this study, we used a p-value threshold of 10^{-4} to select the final change points list ($CP' = (\tau_1, \dots, \tau_{K'})$) that is a subset of the original CP .

Algorithm 1 TV-1D algorithm to remove staircases

```

1: function REMOVESTAIRCASE(R)
2:    $detected_F \leftarrow \emptyset$ 
3:   while  $length(R) > 0$  do
4:      $N \leftarrow length(R)$ ,  $\sigma \leftarrow std(R)$ ,  $\epsilon \leftarrow \sigma \times \sqrt{2 \times \log(N)}$ 
5:      $i = \bar{i} = \underline{i} = \bar{h} = \underline{h} = 0$ ,  $R_{up} = R_0 + \epsilon$ ,  $R_{down} = R_0 - \epsilon$ 
6:     while  $i < N$  do
7:       if  $i < N - 1$  then  $\hat{\epsilon} = \epsilon$ 
8:       else  $\hat{\epsilon} = 0$ 
9:       end if  $\bar{h} = \bar{h} + R_{up} - R_i$ 
10:      if  $\bar{h} > \hat{\epsilon}$  then
11:         $F[i_0 + 1 : \bar{i}] = R_{up}, (i_0, \bar{i}) = \bar{i}, R_{down} = R_i + 2 \times \epsilon$ ,
12:         $R_{up} = R_i, \underline{h} = \epsilon, \bar{h} = -\epsilon, i = \bar{i} + 1$ 
13:        Break
14:      end if
15:      if  $\underline{h} < -\hat{\epsilon}$  then
16:         $F[i_0 + 1 : \underline{i}] = R_{down}, (i_0, \bar{i}) = \underline{i}, R_{up} = R_i - 2 \times \epsilon$ ,
17:         $R_{down} = R_i, \underline{h} = \epsilon, \bar{h} = -\epsilon, i = \underline{i} + 1$ 
18:        Break
19:      end if
20:      if  $\bar{h} \leq -\hat{\epsilon}$  then  $R_{up} = R_{up} + \frac{\hat{\epsilon} - \bar{h}}{i - i_0}$ ,  $\bar{h} = -\hat{\epsilon}$ ,  $\bar{i} = i$ 
21:      end if
22:      if  $\underline{h} \geq \hat{\epsilon}$  then  $R_{down} = R_{down} + \frac{-\hat{\epsilon} - \underline{h}}{i - i_0}$ ,  $\underline{h} = \hat{\epsilon}$ ,  $\underline{i} = i$ 
23:      end if
24:       $i = i + 1$ 
25:    end while
26:     $R = F[i_0 + 1 : N] = R_{up}, detected_F += F[1 : i_0]$ 
27:  end while
28:  Return  $detected_F$ 
29: end function

```

2.1.3 Calling CNV segments.—By calculating the mean of signal $\mathbf{r}$ between each pair of high confidence change points ($\{\tau_1, \dots, \tau_K'\}$), the segmentation method reconstructs the CNV signal and calls CNV segments with their values. Finally, it maps back the detected segments to their corresponding genomic regions.

2.1.4 Preprocessing read-count data.—It has been shown that normalization and denoising read-count data can significantly increase the accuracy of CNV detection [34, 36, 38]. We developed a preprocessing method which consists of two parts to increase the detection accuracy of the proposed segmentation algorithm. We generated a read-count sequence of concatenated exonic regions from a sample, $\mathbf{d}_s$, and control, $\mathbf{d}_c$, read-count data. We calculated the ratios between read-counts of sample and read-counts of control $\mathbf{d}_{s,c}$.

Correcting the imbalance library size effect.: To correct the imbalance library size effect, we multiply $\mathbf{d}_{s,c}$ by the ratio of the sums of read-count of the sample and control in all

exonic regions [11, 15] as $\mathbf{r} = \mathbf{d}_{s,c} \times \frac{\sum_{i=1}^e \sum_{j=1}^{n_i} c_j^i}{\sum_{i=1}^e \sum_{j=1}^{n_i} s_j^i}$, where $\mathbf{r}$ corrected read-count ratio signal

with length N , e is the total number of exonic regions, s_j^i and c_j^i are read-count of the j th genomic position of the i th exonic region for the sample and control, respectively.

Removing outliers.: The second step in the preprocessing of the read-count signal is to identify and remove outliers [2, 21, 22]. We used the Hampel identifier that is robust against

outliers and does not require prior knowledge of the data distribution [16]. It calculates the median in a window that contains r_i and l points on either side of r_i . Given a sequence of data $[r_1, r_2, r_3, \dots, r_N]$ and a sliding window of length l , the method computes point-to-point median and standard-deviation estimates as $m_i = \text{median}(r_{i-l}, r_{i-l+1}, r_{i-l+2}, \dots, r_{i+l-2}, r_{i+l-1}, r_{i+l})$ and $\sigma_i = \kappa \times \text{median}(|r_{i-l} - m_i|, \dots, |r_{i+l} - m_i|)$, where m_i is the local median, σ_i is the standard deviation and $\kappa = \frac{1}{(\sqrt{2} \times \text{erfc}^{-1}(1/2))} = 1.4826$. The factor κ makes an unbiased estimation of the standard deviation for Gaussian data [23]. The quantity σ_i/κ is known as the median absolute deviation (MAD). For a sample r_i and a given threshold n_σ , if $|r_i - m_i| > n_\sigma \sigma_i$, then the Hampel identifier declares r_i as an outlier and replaces it with m_i [16]. In this work, we used the threshold of $n_\sigma = 0.1$ and a sliding window of length $100bp$ for the simulated WES data. Due to the high level of noise in the real data we used a threshold of $n_\sigma = 0.1$ and a sliding window of length $1000bp$ for the real read-count signals.

2.2 Data Sets

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

2.2.1 Simulated read-count data sets.—We generated 100 simulated read-count signals with known true CNV segments as the gold standard. In addition to these read-count data, we generated 100 noisy read-count signals by adding different levels of Gaussian white noise to each of the simulated signals. We used a range of $SNRs$ from 0.1 to 10, where SNR is defined as the ratio of the signal power (meaningful information) to the background noise power (unwanted signal).

2.2.2 Read-count generation from simulated and real WES sequencing data sets.—We used a CNV simulator, called CNV-Sim (<https://github.com/NabaviLab/CNV-Sim>) to simulate WES data with known CNVs. For chromosome 1, we generated ten paired-end WES datasets with the read length of 100 bp. We used the BWA tool [13] to align short reads to the reference genome (hg19) and generated BAM files. Also, we used WES data from nine breast cancer tumor and matched normal. The aligned BAM files of these nine 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 nine tumor samples as a benchmark, downloaded from the TCGA data portal (<https://portal.gdc.cancer.gov/projects/TCGA-BRCA>). For both simulated and real data, to generate a read-count signal, first, we sorted BAM files and removed duplicated reads using SAMTools [14] and Picard[24] tools. Then, we calculated the base-level read-count of the exonic regions of both sample and control data using BEDTools [26].

3 RESULTS AND DISCUSSION

3.1 Results on simulated datasets

The CBS method [19] is a well-established segmentation method and many CNV detection tools use it for identifying CNV segments [9, 15, 17, 31, 33]. We applied CBS to the noisy simulated datasets to compare its performance with that of our segmentation method. We

compared the detected segments with benchmark segments as in [37] to call true positives (TPs), false negatives (FNs), and false positives (FPs). We used the "GenomicRanges" R package from Bioconductor (<http://bioconductor.org/packages/GenomicRanges/>) to calculate overlapping regions between detected CNVs and benchmark CNVs. The threshold of ± 0.2 for $\log_2 \text{ratio}$ and 50% overlap were used for calling CNV segments. In this work, we applied the DNAcopy R package [32] to apply CBS on noisy signals considering different values of undo.SD parameter (ranging from 4 to 12). The undo.SD parameter indicates the number of standard deviations (SDs) between the means of two adjacent regions to have a split. Figure 4 shows the proposed method outperforms CBS in detecting TPs for all undo.SD values, especially for low SNR data where the level of noise is high. The sensitivity of the proposed method is almost 1 for a wide range of SNRs . The proposed method has a comparable performance with CBS in detecting FPs. One of the drawbacks of CBS is its dependency on the user-defined parameters where different parameters lead to different performances.

3.2 Results on simulated and real read-count data

We compared the performance of our proposed method with four popular CNV detection tools: Varscan2 [10], ExomeCNV [31], cn.Mops [9], and Contra [15] using simulated and real sequencing data (chromosome 1). All these tools use CBS for the segmentation. To call TPs, TNs, FPs, and FN, we compared the CNV status of genes as in [37]. We annotated the detected CNV segments to obtain gene lists. We used the "cghMCR" R package from Bioconductor (<http://bioconductor.org/packages/cghMCR/>) to identify CNV genes using Refseq gene identifications. Thresholds of ± 0.2 were used to call CNV genes. The average of sensitivities, false discover rates (FDRs), specificities, F1-score, and accuracy for detecting amplified and deleted genes are calculated and shown in Table 1 and 2. For simulated data, the proposed method provides the highest sensitivity (95.76%, 88%), F1-Score (90.18%, 84.90%), and accuracy (99.85%, 99.87%) in detecting amplified and deleted CNV genes, respectively; and it has comparable FDR and specificity compared to the other tools. The highest sensitivities indicate the ability of our proposed method to discover more true CNV regions. F1-score considers both the precision (1-FDR) and the sensitivity, and the highest value of F1-score shows the ability of our proposed method to detect more TPs and less FPs and FN compared to the other tools. For real data, our developed method outperforms the other tools with the lowest FDR (9.96%) and the highest sensitivity (86.77%), specificity (94.04%), F1-score (88.37%), and accuracy (90.80%) for amplification. For deleted CNVs, the proposed method shows the highest performance in sensitivity. All of the tools have poor performance in detecting deleted CNVs, and our proposed method has comparable performance compared to other tools. We need to mention that we used the CNV results from the array-based technology for benchmarking. The resolution of the array-based technology in detecting CNV is very low compared to the NGS technology. The high FDRs, especially in deletion, can be partially due to the limitation of the array-based technology.

4 RUNTIME COMPARISON

Many CNV detection tools use CBS for calling CNV segments [37]. CBS uses an iterative algorithm based on the variance of the data. One of the main drawbacks of CBS is its

running time, especially for high throughput NGS data. We compared the efficiency of the modified Taut String algorithm with CBS. Using 100 simulated datasets 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 an average 450 seconds to 490 seconds for different values of undo.SD parameter (ranging from 12 to 4). However, it takes on average about 85 seconds for the proposed method. The main reason for the lower overall runtime of our algorithm is due to employing the Taut String algorithm. The complexity of Taut String algorithm is linear in time, and it can improve the efficiency of detection of CNV segments.

5 CONCLUSION

In this study, we showed that treating read-count data as sparse PWC signals and using a new segmentation method based on TV optimization can result in more precise CNV detection. The proposed method that is inspired by the Taut String algorithm detects high confidence change points to reconstruct PWC signals as CNV segments. The original Taut String algorithm is a very efficient approach with linear time complexity to solve the TV optimization problem. However, it suffers from false change point detection due to inappropriately choosing the regularization parameter and the staircase effect. Because the first detected change point of TV is not affected by the staircase effect, the proposed algorithm iteratively detects the first change point of the read-count signal and resets again with a new start point. In each iteration, the regularization parameter is changed based on the length and the standard deviation of a new shorter signal. Therefore at each iteration, the algorithm reduces the staircase effect and at the same time weighs signal by different regularization parameter, which result in more accurate CNV detection. In addition, the developed segmentation method assigns p-values to change points using the Pettitt test and filters out low confidence change points using a threshold that results in lower FDRs in detecting CNV segments. Our proposed method outperforms existing CNV detection tools in terms of accuracy and FDR. In summary, the proposed approach is an efficient segmentation method, and it provides very high sensitivities in detecting CNVs.

REFERENCES

- [1]. Barbero Á Ivaro and Sra Suvrit. 2014 Modular proximal optimization for multidimensional total-variation regularization. arXiv:1411.0589 [math, stat] (Nov. 2014). <http://arxiv.org/abs/1411.0589> arXiv: 1411.0589.
- [2]. Alvaro Furlani Bastos Keng-Weng Lao, Todeschini Grazia, and Santoso Surya. 2018 Novel Moving Average Filter for Detecting RMS Voltage Step Changes in Triggerless PQ Data. IEEE Transactions on Power Delivery 33, 6 (Dec. 2018), 2920–2929. 10.1109/TPWRD.2018.2831183
- [3]. Pierre Raphael Bertrand Mehdi Fhima, and Guillin Arnaud. 2011 Off-Line Detection of Multiple Change Points by the Filtered Derivative with p -Value Method. Sequential Analysis 30, 2 (4 2011), 172–207. 10.1080/07474946.2011.563710
- [4]. Cho Haeran and Fryzlewicz Piotr. 2011 Multiscale interpretation of taut string estimation and its connection to Unbalanced Haar wavelets. Statistics and Computing 21, 4 (Oct. 2011), 671–681. 10.1007/s11222-010-9200-5
- [5]. Lutz DÄijmbgen Arne Kovac, and others. 2009 Extensions of smoothing via taut strings. Electronic Journal of Statistics 3 (2009), 41–75.
- [6]. Duan Junbo, Zhang Ji-Gang, Deng Hong-Wen, and Wang Yu-Ping. 2013 CNV-TV: A robust method to discover copy number variation from short sequencing reads. BMC Bioinformatics 14, 1 (2013), 150 10.1186/1471-2105-14-150 [PubMed: 23634703]

- [7]. Fridlyand Jane, Snijders Antoine M., Pinkel Dan, Albertson Donna G., and Jain Ajay N.. 2004 Hidden Markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis* 90, 1 (7 2004), 132–153. 10.1016/j.jmva.2004.02.008
- [8]. Hagen Lene. 2017 Lasso-Path and Taut String Algorithm for One-Dimensional Total Variation Regularization. (2017), 58 <https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2459271>
- [9]. Klambauer GÄijnter, Schwarzbauer Karin, Mayr Andreas, Clevert Djork-ArnÄ , Mitterecker Andreas, Bodenhofer Ulrich, and Hochreiter Sepp. 2012 cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Research* 40, 9 (5 2012), e69–e69. 10.1093/nar/gks003 [PubMed: 22302147]
- [10]. Koboldt Daniel C., Zhang Qunyan, Larson David E., Shen Dong, McLellan Michael D, Lin Ling, Miller Christopher A., Mardis Elaine R., Ding Li, and Wilson Richard K.. 2012 VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* 22, 3 (3 2012), 568–576. 10.1101/gr.129684.111 [PubMed: 22300766]
- [11]. Kong Jinhwa, Shin Jaemoon, Won Jungim, Lee Keonbae, Lee Unjoo, and Yoon Jeehee. 2017 ExCNVSS: A Noise-Robust Method for Copy Number Variation Detection in Whole Exome Sequencing Data. *BioMed Research International* 2017 (2017), 1–11. 10.1155/2017/9631282
- [12]. Kulkarni Pranav and Frommolt Peter. 2017 Challenges in the Setup of Large-scale Next-Generation Sequencing Analysis Workflows. *Computational and Structural Biotechnology Journal* 15 (2017), 471–477. 10.1016/j.csbj.2017.10.001 [PubMed: 29158876]
- [13]. Li H and Durbin R. 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 14 (7 2009), 1754–1760. 10.1093/bioinformatics/btp324 [PubMed: 19451168]
- [14]. Li Heng, Handsaker Bob, Wysoker Alec, Fennell Tim, Ruan Yue, Homer Nils, Marth Gabor, Abecasis Goncalo, Durbin Richard, and others. 2009 The sequence alignment/map format and SAMtools. (2009), 2078–2079 pages.
- [15]. Li Jason, Lupat Richard, Amarasinghe Kaushalya C., Thompson Ella R., Doyle Maria A., Ryland Georgina L., Tothill Richard W., Halgamuge Saman K., Campbell Ian G., and Gorringer Kylie L.. 2012 CONTRA: copy number analysis for targeted resequencing. *Bioinformatics* 28, 10 (5 2012), 1307–1313. 10.1093/bioinformatics/bts146 [PubMed: 22474122]
- [16]. Liu Hancong, Shah Sirish, and Jiang Wei. 2004 On-line outlier detection and data cleaning. *Computers & Chemical Engineering* 28, 9 (Aug. 2004), 1635–1647. 10.1016/j.compchemeng.2004.01.009
- [17]. Magi Alberto, Tattini Lorenzo, Cifola Ingrid, Romina Dâ Aurizio Matteo Benelli, Mangano Eleonora, Battaglia Cristina, Bonora Elena, Kurg Ants, Seri Marco, Magini Pamela, Giusti Betti, Romeo Giovanni, Pippucci Tommaso, Gianluca De Bellis Rosanna Abbate, and Gian Franco Gensini. 2013 EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biology* 14, 10 (2013), R120 10.1186/gb-2013-14-10-r120 [PubMed: 24172663]
- [18]. Mallakpour Iman and Villarini Gabriele. 2016 A simulation study to examine the sensitivity of the Pettitt test to detect abrupt changes in mean. *Hydrological Sciences Journal* 61, 2 (Jan. 2016), 245–254. 10.1080/02626667.2015.1008482
- [19]. Olshen Adam B, Venkatraman ES, Lucito Robert, and Wigler Michael. 2004 Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5, 4 (2004), 557–572. [PubMed: 15475419]
- [20]. Ottersten Johan, Wahlberg Bo, and Rojas Cristian. 2016 Accurate Changing Point Detection for 11 Mean Filtering. *IEEE Signal Processing Letters* (2016), 1–1. 10.1109/LSP.2016.2517605
- [21]. Pearson RK. 2001 Exploring process data. *Journal of Process Control* 11, 2 (4 2001), 179–194. 10.1016/S0959-1524(00)00046-9
- [22]. Pearson RK. 2002 Outliers in process modeling and identification. *IEEE Transactions on Control Systems Technology* 10, 1 (Jan. 2002), 55–63. 10.1109/87.974338
- [23]. Pearson Ronald K., Neuvo YrjÄ , Astola Jaakko, and Gabbouj Moncef. 2016 Generalized Hampel Filters. *EURASIP Journal on Advances in Signal Processing* 2016, 1 (Dec. 2016), 87 10.1186/s13634-016-0383-6

- [24]. Picard Franck, Robin Stephane, Lavielle Marc, Vaisse Christian, and Daudin Jean-Jacques. 2005 A statistical approach for array CGH data analysis. *BMC bioinformatics* 6 (Feb. 2005), 27 10.1186/1471-2105-6-27 [PubMed: 15705208]
- [25]. Pohlert Thorsten. Non-parametric trend tests and change-point detection. (????).
- [26]. Quinlan Aaron Rand Hall Ira M. 2010 BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 6 (2010), 841–842. [PubMed: 20110278]
- [27]. Genetics Home Reference. 2019 What are whole exome sequencing and whole genome sequencing? (2019). <https://ghr.nlm.nih.gov/primer/testing/sequencing>
- [28]. Roca Iria, Lorena Gonzlez-Castro Helena Fernndez, Couce Mluz, and Fernndez-Marmiesse Ana. 2019 Free-access copy-number variant detection tools for targeted next-generation sequencing data. *Mutation Research/Reviews in Mutation Research* 779 (Jan. 2019), 114–125. 10.1016/j.mrrev.2019.02.005
- [29]. Rojas Cristian R. and Wahlberg Bo. 2014 On change point detection using the fused lasso method. arXiv:1401.5408 [math, stat] (Jan. 2014). <http://arxiv.org/abs/1401.5408> arXiv: 1401.5408.
- [30]. Rojas Cristian R. and Wahlberg Bo. 2015 How to monitor and mitigate staircasing in L1 trend filtering In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, South Brisbane, Queensland, Australia, 3946–3950. 10.1109/ICASSP.2015.7178711
- [31]. Jarupon Fah Sathirapongsasuti Hane Lee, Horst Basil A. J., Brunner Georg, Cochran Alistair J., Binder Scott, Quackenbush John, and Nelson Stanley F. 2011 Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* (Oxford, England) 27, 19 (Oct. 2011), 2648–2654. 10.1093/bioinformatics/btr462
- [32]. Olshen Venkatraman Adam Seshan E. 2017 DNACopy. (2017). 10.18129/B9.bioc.DNACopy
- [33]. Wang Chen, Evans Jared M, Bhagwate Aditya V, Prodduturi Naresh, Sarangi Vivekananda, Middha Mridu, Sicotte Hugues, Vedell Peter T, Hart Steven N, Oliver Gavin R, and others. 2014 PatternCNV: a versatile tool for detecting copy number changes from exome sequencing data. *Bioinformatics* 30, 18 (2014), 2678–2680. [PubMed: 24876377]
- [34]. Zare Fatima, Ansari Sardar, Najarian Kayvan, and Nabavi Sheida. 2017 Noise cancellation for robust copy number variation detection using next generation sequencing data. *IEEE*, 230–236. 10.1109/BIBM.2017.8217654
- [35]. Zare Fatima, Ansari Sardar, Najarian Kayvan, and Nabavi Sheida. 2018 Copy number variation detection using partial alignment information In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, Madrid, Spain, 2435–2441. 10.1109/BIBM.2018.8621529
- [36]. Zare Fatima, Ansari Sardar, Najarian Kayvan, and Nabavi Sheida. 2018 Preprocessing Sequence Coverage Data for Precise Detection of Copy Number Variations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2018), 1–1. 10.1109/TCBB.2018.2869738 [PubMed: 27542180]
- [37]. Zare Fatima, Dow Michelle, Monteleone Nicholas, Hosny Abdelrahman, and Nabavi Sheida. 2017 An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC bioinformatics* 18, 1 (2017), 286. [PubMed: 28569140]
- [38]. Zare Fatima, Hosny Abdelrahman, and Nabavi Sheida. 2018 Noise cancellation using total variation for copy number variation detection. *BMC Bioinformatics* 19, S11 (Oct. 2018). 10.1186/s12859-018-2332-x

Figure 1:
The overall block diagram of the proposed method

Figure 2:

Staircase effect of Taut String for a) SNR=5, b) SNR=15; and the importance of regularization parameter for SNR=15 when c) ϵ is a low value and when d) ϵ is a high value.

Red dots represent the observed noisy data, green lines represent desired signal and black lines represent detected signal.

Figure 3:
The performance of the proposed algorithm in removing staircase effects.

Figure 4:

The performance of the proposed method and CBS with different values for `undo.SD` parameter in detecting CNV segments in terms of a) sensitivity b) FDR.

Table 1: Overall Performance of CNV Detection Methods Using the Simulated WES data (Chromosome 1)

Method	Amplification					Deletion				
	Sensitivity	FDR	Specificity	F1-score	Accuracy	Sensitivity	FDR	Specificity	F1-score	Accuracy
Our proposed method	95.76%	14.78%	99.88%	90.18%	99.85%	88%	17.99%	99.85%	84.90%	99.78%
VarScan2	82.98%	7.98%	99.94%	87.26%	99.82%	80.55%	17.71%	99.89%	81.41%	99.77%
ExomeCNV	76.90%	4.87%	99.97%	85.04%	99.83%	74.19%	6.87%	99.96%	82.58%	99.80%
cn.Mops	78.40%	17.83%	99.87%	80.23%	99.72%	56.07%	10.28%	99.94%	69.02%	99.67%
Contra	89.73%	21.85%	99.81%	83.54%	99.74%	96.10%	32.15%	99.69%	79.53%	99.67%

Overall Performance of CNV Detection Methods Using Real Data (Chromosome 1)

Table 2:

Method	Amplification					Deletion				
	Sensitivity	FDR	Specificity	F1-score	Accuracy	Sensitivity	FDR	Specificity	F1-score	Accuracy
Our proposed method	86.77%	9.96%	94.04%	88.37%	90.80%	81.47%	60.54%	82.44%	53.16%	83.28%
VarScan2	72.20%	34.18%	90.82%	68.86%	86.14%	75.34%	43.27%	56.72%	64.72%	79.67%
ExomeCNV	85.30%	44.91%	78.81%	66.94%	78.89%	80.90%	50.77%	85.97%	61.21%	85.53%
cn.Mops	55.56%	61.62%	65.84%	45.39%	65.15%	52.53%	65.21%	76.43%	41.85%	72.76%
Contra	58.93%	58.17%	81.59%	48.92%	75.39%	64.70%	68.37%	77.39%	42.48%	74.25%