33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Detection of Copy Number Variation using Scale Space Filtering

Jongkeun Lee, Baeksop Kim, Jeehee Yoon, and Unjoo Lee

Abstract— **This study proposes a novel CNV detection algorithm based on scale space filtering. It uses Gaussian filter for the convolution with a scale parameter. The range of the scale parameter is adjusted according to the coverage level of read data. The position of a CNV region is determined through a coarse and a fine searches over the scales. The results showed low dependency of the performance of the proposed method on the coverage level compared to the conventional methods. The results also showed that the proposed method outperforms the conventional methods by 63.29 ~ 73.57 %.**

I. Introduction

COPY number variation (CNV), a form of structural variants in human genome is an event in which a large DNA fragment (> 1 kbp) has population differences by duplications or deletions. It has been known to be associated with human genetic diseases such as Crohn's disease and type 1 diabetes, and linked to a range of disorders including schizophrenia, autism, and birth defect syndromes [1].

There are two typical ways of detecting CNVs: one is to use microarray-based methods [2]-[3] and the other is to use sequence-based methods [4]-[5]. A major drawback of the array-based CNV detection method is the low resolution depending on the array technology. Sequence-based detection of CNVs can be approached in two ways. One way is to directly compare accurately completed sequence assemblies of the genomes; the other way to use giga-sequencing data, such as paired-end reads generated by next-generation sequencing technology. The latter is more cost-effective than the former. Recently, tools for CNV detection with giga-sequencing data were developed by using depth of read coverage [5]-[7] and paired-end read mapping (PEM) [8]. The method based on PEM has difficulties in detecting CNVs in a region with complex structural variants since it is dependent on the technology of the paired-end read generation. The method using depth of read coverage generally requires high levels of read coverage [5]. However, CNV-seq [6], using the ratios of reference and test read coverage, enables the detection of CNV regions at relatively low levels of read coverage. CNV_shape [7] also makes it possible to detect CNV regions at low levels of read coverage through the Gaussian normalization and shape-based extraction algorithm.

While the advent of the next-generation sequence technology makes it available more detailed and individualized analysis of human genome, CNV detection methods still need an improvement in the precisions of the positions and the sizes which are essential for personalized prediction of a specific disease or a genetic deficiency.

This study proposes a novel CNV detection algorithm based on scale space filtering. The scale space filtering is a method that describes one or two dimensional signal through generating successively higher level descriptions of the signal by convolving it with a filter [9]. The proposed method uses Gaussian filter for the convolution with a scale parameter, the range of which is adjusted according to the level of read coverage. As a result, it is possible to detect the exact positions of various size of CNVs from giga-sequencing data with relatively low-level of read coverage. The proposed method was verified by experiments using simulated and real data. The results showed that the proposed method has low dependency of its performance on the level of the read coverage compared to the conventional methods

II. Method

Figure 1 presents a functional block diagram of the overall process of the proposed method. The proposed method consists of two stages. In the first stage, the scale space filtering of input data *C(i)* is obtained by convolution with a Gaussian mask $g(j,\sigma)$ over a continuum of sizes and then inflection points of the scale space image $C(i,\sigma)$ at all values of σ are given through the Laplacian and zero crossing detection. In the second stage, CNV regions are detected using interval means obtained from the inflection points.

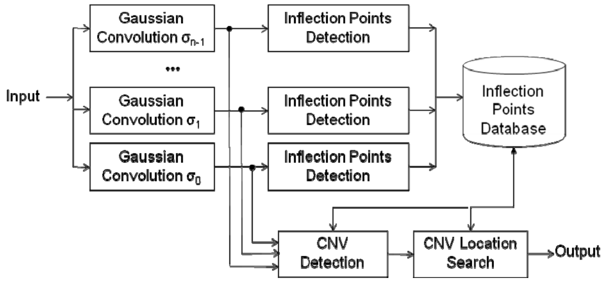
Input
Gaussian Convolution σ_{n-1}
•••
Gaussian Convolution σ_1
Gaussian Convolution σ_0
Inflection Points Detection
Inflection Points Detection
Inflection Points Detection
Inflection Points Database
CNV Detection
CNV Location Search
Output

Fig. 1. Block diagram of the proposed method

Manuscript received March 26, 2011. This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0003706).

Jongkeun Lee (e-mail: jongkeun@hallym.ac.kr), Baeksop Kim (e-mail: bskim@hallym.ac.kr), and Jeehee Yoon (corresponding author, phone:+82-33-248-2316; fax:+82-33-242-2524; e-mail: jhyoon@hallym.ac.kr) are with the Department of Computer Engineering, Hallym University, Korea.

Unjoo Lee is with the Department of Electronic Engineering, Hallym University, Korea (e-mail: ejlee@hallym.ac.kr).

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

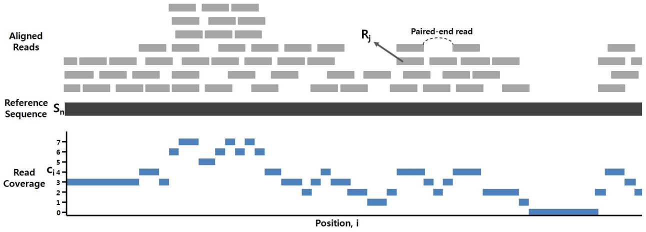

Fig. 2. An overview diagram of the processes of obtaining read coverage data

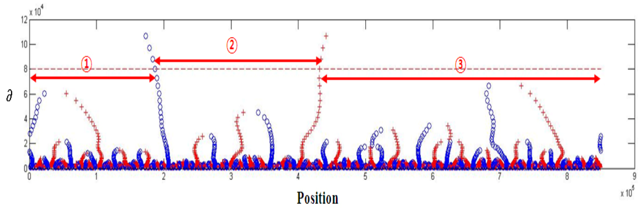

Fig. 4. Zero contours of $\nabla_i \cdot C = 0$ in a scale-space image of a read coverage data *C(i)*.

A. Input Data : Read coverage data is used as input data in our proposed method. Given a standard reference DNA sequence *S* and the read sequence R_j of a given test sequence *T*, read coverage data *C(i)* of *T* is obtained by summing up the numbers of read sequence aligned to the position *i*. Figure 2 shows an overview diagram of the processes of obtaining read coverage data.

B. Scale Space Filtering : The Gaussian convolution $C(i,\sigma)$ of read coverage data *C(i)* is obtained by

$$C(i,\sigma) = C(i) * g(j,\sigma) = \sum_{j=-n}^{n} c_{i-j} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{j^2}{2\sigma^2}}$$

where σ is the smoothing parameter considered with 1.1 of the ratio of the successive scales in the range from 1000 to 10^6, n is set to $\lceil 3\sigma \rceil$ and the Gaussian function satisfies $\sum_{j=-n}^{n} g(j,\sigma) = 1$. Figure 3 shows the graph of the scale-space image $C(i,\sigma)$ of a read coverage data *C(i)* with increasing σ .

C. Inflection points detection : In terms of the scale-space image $C(i,\sigma)$, inflection points at all values of σ are the points that satisfy $\nabla_i \cdot C(i,\sigma) = 0$. The contours of $\nabla_i \cdot C(i,\sigma) = 0$ mark the appearance and trajectory of inflection points in the smoothed signal. Figure 4 shows zero contours of $\nabla_i \cdot C(i,\sigma) = 0$ in a scale-space image $C(i,\sigma)$ of a read coverage data *C(i)*. Notice that the zero contours form arches, closed above, but open below. Each arch consists of a pair of contours, crossing zero with opposite sign. Therefore, an interval between two adjacent zero contours at a certain value of σ can be a candidate CNV region. Some of these intervals are plotted with horizontal lines with arrows in Figure 4. The positions of zero contours at all values of σ are stored in the inflection point database.

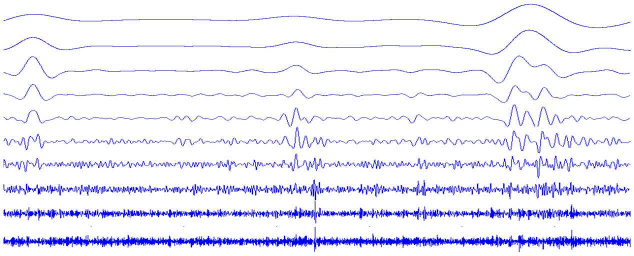

Fig. 3. Scale-space image $C(i,\sigma)$ of a read coverage data *C(i)* with increasing σ .

D. Detection of CNV regions at a coarse search : The mean of each interval of the scale-space image $C(i,\sigma)$ is calculated as follows:

$$m_I(k,\sigma) = \frac{1}{|I_k|} \sum_{i \in I_C(k,\sigma)} C(i,\sigma)$$

where $|I_k|$ is the size of the k-th interval $I_C(k,\sigma)$ of the scale-space image $C(i,\sigma)$ at a given σ . The interval $I_C(k,\sigma)$ is called as a CNV gain or loss region when it satisfies the conditions $m_I(k,\sigma) \geq 1.5 m_C(\sigma)$ or $m_I(k,\sigma) \leq 0.5 m_C(\sigma)$, respectively, where $m_C(\sigma)$ is the mean of the entire region of the scale-space image $C(i,\sigma)$ at a given σ .

E. *Determination of CNV locations at a fine* search : Once a CNV region is detected at a coarse scale, then its exact location is determined at a fine scale. A neighbor of an inflection point at a given σ is the nearest inflection point at the next finer scale $\sigma/1.1$. The sequence of neighbors constitutes the path from an inflection point at a coarse scale to the corresponding points at the finer scales. Therefore, the inflection points at a coarse scale are reduced to the corresponding points at the finer scale as shown in Figure 4. The last point of the path, that is, the point at the finest scale becomes the exact location of the first point of the path, that is, the point at the coarse scale. Therefore, the start and end positions of the CNV detected at a coarse scale is determined to the corresponding inflection points of the scale-space image $C(i,\sigma)$ at the finest scale.

III. Results and Discussion

A. Experiment with simulated data

A simulation data generator was constructed to generate data for simulation. The simulation data generator initializes a given DNA sequence as a reference sequence and a test sequence. It then copies some of the corresponding CNV regions reported in the CNV database of the DGV [10] and puts them in random positions of the reference sequence or test sequence so that the test sequence has CNV gain or loss regions that differ in size and location from those of the reference sequence. Once the reference sequence and test

TABLE I.
THE ASSESSMENT OF THE PROPOSED METHOD FOR VARIOUS READ COVERAGE LEVELS (E=3%) AND FOR VARIOUS SEQUENCE ERROR RATES (C=3×)

Read dept level (C)	FPR	FNR	Sequence error rate (E)	FPR	FNR
0.5×	0.450	6.526	1%	0.144	0.901
1×	0.251	4.543	2%	0.155	1.376
3×	0.260	3.435	3%	0.260	3.435
6×	0.169	1.356	4%	0.382	17.514
10×	0.215	1.584	5%	0.570	37.472

sequence are generated, the test sequence is treated to have a given error sequence rate and then reads of the test sequence are generated by Solexa machine.

NCBI Build 36.3 chromosome 8 genomic contig NT_077531.3 was used for the generation of the simulated data. SOAP (Short Oligonucleotide Alignment Program) [11] was used for the alignment of the read data and a random match method, one of various alignment algorithms that SOAP supports was used with e = 2 mismatch criteria as a tolerable limit with regard to noise, such as sequence errors. The experiments were conducted in the platforms of Windows 7 and CentOS 5.5 with an Intel Core i7 2.8 GHz CPU, 8 GB of main memory, and a 2 TB hard drive. The programming language used for the development of the proposed method was MATLAB.

The experiments were carried out for the assessment of the proposed method for various read coverage levels. The read coverage levels used were 0.5×, 1×, 3×, 6×, and 10×. More than 20 experiments for each read coverage level were accomplished with different simulated data, the results from which were averaged for the assessment. Here, the sequence error rate $E = 3\%$ was considered according to a typical error rate existing in real data generated by next sequencing technology, even though the sequence error rate keeps improving due to the advancement of the technology.

The other experiments were carried out for the assessment of the proposed method for various sequence error rates. The values 1%, 2%, 3%, 4%, and 5% of the sequence error rate were used. The level of read coverage used here was $C = 3\times$.

Table I summarizes the assessment results for various read depth levels and for various sequence error rates. As shown in Table I, the proposed method has fairly good values of FPR and FNR at even very low-level of read coverage $C = 0.5\times$ and the performances increase as the level of read coverage increases and the sequence error rate decreases. The abrupt change of FNR at the sequence error rate $\mathrm{E} \geq 4\%$ seems to be caused by the increase of the failure rate ($\geq 20 \sim 30\%$) of read alignment.

B. *Experiment with real human data*

Paired-end read data of NA18507 and NA10851 downloaded from the site of the 1000 Genomes project (*http://www.1000genomes.org*) were used for the experiments with real human data; they were generated by the Solexa GA machine. The average coverage level of the downloaded data were 1.7× for NA18507 and 5.6× for NA10851. The performance of the proposed method was

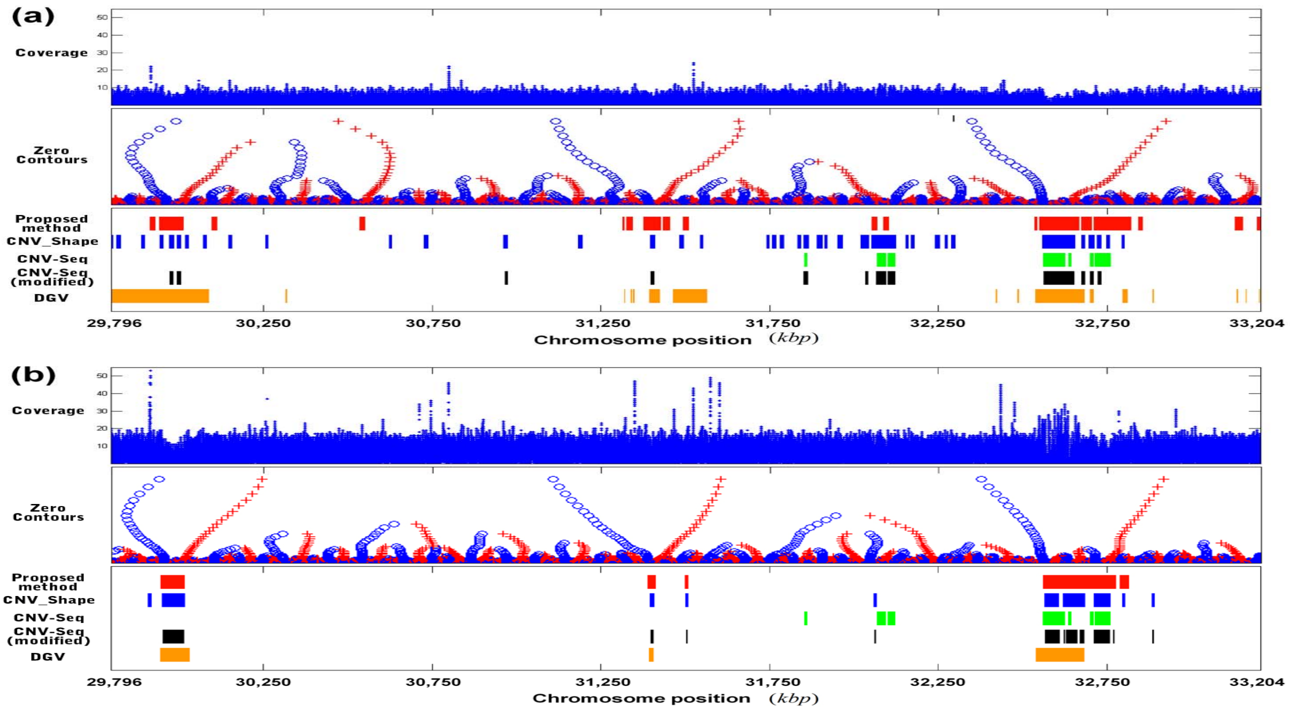

Fig. 5. Experimental results of the human leukocyte antigen (HLA) region of chr. 6 of NA18507 (Figure 5(a)) and NA10851 (Figure 5(b)).

assessed by estimating the FNR and the FPR on the basis of the CNV database of the DGV and comparing them to the corresponding values of CNV-seq [6], the modified CNV-seq [6], and CNV_shape [7]. The modified CNV-seq is an implementation of CNV-seq modified for detecting CNV regions individually without using relative differences between read coverage data.

Figure 5 shows the experimental results of the human leukocyte antigen (HLA) region of chr. 6 of NA18507 (Figure 5(a)) and NA10851 (Figure 5(b)). The HLA region resides on the short arm of human chr. 6 and is 3.408 Mbp long; it ontains around 200 genes related to immune system function in humans. The top panels of Figures 5(a) and 5(b) show graphs of the read coverage data of NA18507 and NA10851, respectively; the x-axis is the position of chr. 6, and the y-axis is the number of reads mapped to each position. The middle panels of Figures 5(a) and 5(b) display plots of the zero contours of the read coverage data of NA18507 and NA10851, respectively; the x-axis is the position of chr. 6, and the y-axis is the values of σ . The bottom panels of Figures 5(a) and 5(b) show the CNV regions detected on NA18507 and NA10851, respectively, by the proposed method along with the CNV regions reported in the CNV database of the DGV; they also show the CNV regions detected by CNV-seq, modified CNV-seq, and CNV_shape for comparisons. As observed in the middle and bottom panels of Figure 5, the regions with relatively high (low) values of read coverage data in the graphs of the middle panels are mapped to the CNV gain (loss) regions, detected by the proposed method as in the bottom panels, confirming the intuitive and reliable nature of the method, which is based on the scale space filtering and the interval mean calculation of the read coverage data. This method accurately detects the CNV gain and loss regions for both NA18507 and NA10851. Furthermore, as shown in the plot (Figure 5(a)) of CNV regions detected by this method, many small regions are detected as CNV regions on NA18507 with a relatively low level of coverage (1.7×).

On the other hand, the CNV regions detected by CNV-seq on NA18507 and NA10851 fail to include regions in which shape variations in the coverage data of the test sequence are the same as those of the control sequence. Table II gives a comparative summary of the performance of the proposed method, CNV_shape, the conventional CNV-seq, and the modified CNV-seq on HLA regions of human chr. 6 of NA18507 and NA10851; the FNR and FPR values were derived on the basis of the CNV database of the DGV for the performance assessment. FNR values of 23.09 % and 18.41 % were derived for NA18507 and NA10851 in this method. In contrast, CNV_shape yielded FNR values of 62.91 % and 26.40 %. The CNV-seq and the modified CNV-seq yielded FNR values of 87.37 % and 77.63 % for NA18507 and 71.35 % and 39.01 % for NA10851, respectively. It appears that the overall FNR and FPR values of NA10851 (5.6×) are lower than those of NA18507 (1.7×), which indicates that the performance of the CNV detection algorithms depends on the coverage level of the read coverage. However, the proposed method has low dependency of its performance on the level of the read coverage compared to the other methods. The results also show that the proposed method outperforms the other methods by 63.29 ~ 73.57 %.

TABLE II.
COMPARATIVE SUMMARY OF THE PERFORMANCE OF THE PROPOSED METHOD, CNV_SHAPE, THE CONVENTIONAL CNV-SEQ, AND THE MODIFIED CNV-SEQ

	NA18507		NA10851	
	FPR	FNR	FPR	FNR
Proposed method	7.95	23.09	2.78	18.41
CNV_shape	11.47	62.90	2.73	26.40
CNV-seq	3.63	87.37	3.48	71.35
CNV-seq(modified)	3.07	77.07	1.78	39.01

IV. Conclusion

This study proposes a novel CNV detection algorithm based on scale space filtering. The proposed method uses Gaussian filter for the convolution with a scale parameter, the range of which is adjusted according to the level of read coverage. It uses a coarse and a fine scale searches for the exact locations of CNV regions. The performance was verified with simulated and real human data, and compared to the conventional methods. It showed improvement in the detection of the positions of various size of CNVs from giga-sequencing data with relatively low-level of read coverage.

References

[1] J. Simpson et al., "Copy number variant detection in inbred strains from short read sequence data," *Bioinformatics*, vol. 26, no. 4, pp. 565–567, 2010.

[2] E. Tuzun et al., "Fine-scale structural variation of the human genome," *Nature Genetics*, Vol. 37, No. 7, pp. 727-732, 2005.

[3] S. A. McCarroll et al., "Integrated detection and population-genetic analysis of SNPs and copy number variation," *Nature Genetics*, Vol. 40, No. 10, pp. 1166-1174, 2008.

[4] R. Khaja et al., "Genome assembly comparison identifies structural variants in the human genome," *Nature Genetics*, Vol. 38, No. 12, pp. 1413-1418, 2006.

[5] A. Abyzon, 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*, 2011.

[6] C. Xie and M. Tammi, "CNV-seq, a new method to detect copy number variation using high-throughput sequencing," *BMC Bioinformatics*, Vol. 10, No. 1, Mar. 2009.

[7] S. K. Hong et al., "Shape-based retrieval of CNV Regions in Read Coverage Data," *InCob 2010*, Sep. 2010.

[8] J. Korbel et al., "PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data," *Genome Biology*, vol. 10, no. 2, R23, Feb. 2009.

[9] A. P. Witkin, "Readings in computer vision : issues, problems, principles, and paradigms," pp. 329-332, 1987.

[10] DGV, Available: http://projects.tcag.ca/variation

[11] R. Li et al.,"SOAP2: an improved ultrafast tool for short read alignment," *Bioinformatics*, Vol. 25, No. 15, pp. 196 6-1967, 2009.