Shape-based retrieval of CNV regions in read coverage data

Sangkyun Hong and Jeehee Yoon*

Department of Computer Engineering,
Hallym University 39 Hallymdaehak-gil, Chuncheon-si,
Kangwon-do 200 702, Republic of Korea
E-mail: kyoons@hallym.ac.kr
E-mail: jhyoon@hallym.ac.kr
*Corresponding author

Dongwan Hong

Cancer Genomics Branch,
National Cancer Center 323 Ilsan-ro, Ilsandong-gu,
Goyang-si, Gyeonggi-do 410 769, Republic of Korea
E-mail: dwhong@ncc.re.kr

Unjoo Lee

Department of Electronic Engineering,
Hallym University 39 Hallymdaehak-gil, Chuncheon-si,
Kangwon-do 200 702, Republic of Korea
E-mail: ejlee@hallym.ac.kr

Baeksop Kim

Department of Computer Engineering,
Hallym University 39 Hallymdaehak-gil, Chuncheon-si,
Kangwon-do 200 702, Republic of Korea
E-mail: bskim@hallym.ac.kr

Sanghyun Park

Department of Computer Science,
Yonsei University 134 Shinchon-dong,
Seodaemun-gu, Seoul, 120 749, Republic of Korea
E-mail: sanghyun@cs.yonsei.ac.kr

Abstract: This study proposes a novel copy number variation (CNV) detection method, CNV_shape, based on variations in the shape of the read coverage data which are obtained from millions of short reads aligned to a reference sequence. The proposed method carries out two transforms, mean shift transform and mean slope transform, to extract the shape of a CNV more precisely from real human data, which are vulnerable to experimental and biological noises. The mean shift transform is a procedure for gaining a preliminary estimation of the CNVs

by statistically evaluating moving averages of given read coverage data. The mean slope transform extracts candidate CNVs by filtering out non-stationary sub-regions from each of the primary CNVs pre-estimated in the mean shift procedure. Each of the candidate CNVs is merged with neighbours depending on the merging score to be finally identified as a putative CNV, where the merging score is estimated by the ratio of the positions with non-zero values of the mean shift transform to the total length of the region including two neighbouring candidate CNVs and the interval between them. The proposed CNV detection method was validated experimentally with simulated data and real human data. The simulated data with coverage in the range of 1× to 10× were generated for various sampling sizes and p-values. Five individual human genomes were used as real human data. The results show that relatively small CNVs (>1 kbp) can be detected from low coverage (> 1.7×) data. The results also reveal that, in contrast to conventional methods, performance improvement from 8.18 to 87.90% was achieved in CNV_shape. The outcomes suggest that the proposed method is very effective in reducing noises inherent in real data as well as in detecting CNVs of various sizes and types.

Keywords: CNV; copy number variation; next-generation sequencing technology; shape-based retrieval.

Reference to this paper should be made as follows: Hong, S., Yoon, J., Hong, D., Lee, U., Kim, B. and Park, S. (2104) 'Shape-based retrieval of CNV regions in read coverage data', *Int. J. Data Mining and Bioinformatics*, Vol. 9, No. 3, pp.254–276.

Biographical notes: Sangkyun Hong received his BS and MS in Computer Engineering in 2005 and 2007, respectively, from Hallym University, Korea. Currently, He is a PhD student at the Department of Computer Engineering, Hallym University, Korea. His research interests include Bioinformatics and Database.

Jeehee Yoon received her MS and PhD in Information Engineering from Kyushu University, Japan in 1985 and 1988, respectively. Currently, she is a Professor at the Department of Computer Engineering, Hallym University. Her research interests include DNA Sequence search, Shape-based retrieval in time-series databases and Microarray data analysis.

Dongwan Hong received his PhD in Computer Engineering from Hallym University in 2008. During 2008–2011, he worked as a senior researcher in Seoul National University, Medical Research Center. He has been a principal investigator at the National Cancer Center, Korea. His research interests include Cancer genomics, RNA-seq, Epigenome, Pathway signalling of cancer and Single molecule sequencing.

Unjoo Lee received her PhD in Electrical Engineering in 1995 from the University of Maryland, College Park, USA. She is an Assistant Professor in the Department of Electronic Engineering, Hallym University, Korea. Her research interests include neuronal signal processing associated with cognition, intention and behaviour and computational bioinformatics, including genomic sequence analysis and synthesis.

Baeksop Kim received his PhD in Electronic Engineering in 1985 from the Korea Advanced Institute of Science and Technology, Seoul. He is currently a Professor

in the Department of Computer Engineering, Hallym University, Chuncheon, Korea. His research interests include pattern recognition, computer vision and artificial intelligence.

Sanghyun Park received his BS and MS in Computer Engineering from Seoul National University in 1989 and 1991, respectively. He received his PhD in Computer Science from UCLA University in 2001. Currently, he is a Professor at the Department of Computer Science, Yonsei University. His research interests include Databases, Data mining and Bioinformatics.

1 Introduction

Structural variants (SVs) in the human genome play an important role in the phenotypic diversity and the genetics of complex diseases (Redon et al., 2006). Personal genome sequencing in the coming years will make it possible to predict a specific diagnosis or inherited genetic losses of a person by detecting the associated variants of the person's genome.

Ever since the human HapMap project was initiated in 2001, single-nucleotide polymorphisms (SNPs) have been used to investigate SVs in the human genome (International HapMap, 2003; Jun et al., 2011). SNP is a DNA sequence transformation or variation that occurs when a single nucleotide in the genome differs among the members of a species. However, recent reports have indicated that SVs can occur on many different scales, from a single base pair (bp) as in SNPs, to more than a few hundred kilo base pairs (kbp) as in insertions, deletions, inversions and copy number variations (CNVs) (Antonina and Bhubaneswar, 2010; Iafrate et al., 2004; Tuzun et al., 2005). Insertions and deletions are caused by the addition and the removal of one or more extra nucleotides, respectively. Inversions are caused by reversing the orientation of a chromosomal segment in the DNA. CNVs, which are known to be very difficult to detect due to sequencing errors, are segments of DNA for which copy number differences can be found by comparing two or more genomes. The sizes of CNVs may range from one kbp to several mega base pairs (Mbp).

In this study, we propose a novel CNV detection method, CNV_shape, which is based on variations in the shape of read coverage data obtained by aligning millions of short reads to a reference sequence. The CNV_shape method carries out two transforms, mean shift transform and mean slope transform, as well as a merging process for more precise modelling of the shape of a CNV event, such as a gain or a loss. The mean shift transform is a procedure for making a preliminary estimate of the CNVs by statistically evaluating the local distribution of a given read coverage data. The mean slope transform extracts candidate CNVs (CCNVs) by filtering out non-stationary sub-regions from each of the primary CNVs pre-estimated in the mean shift transform. Each of the resulting CCNVs is merged with neighbours to be finally identified as a putative CNV.

The CNV_shape method suggested in this study is simple and intuitive. It resolves the problems of sequencing errors inherent in a reference sequence and reads. To maximise the efficiency of CNV_shape, we applied random match, an optimal sequence alignment method for CNV detection, in accordance with the results of our previous work (Lee et al., 2009).

We carried out simulation experiments for various window sizes (0.1–8 kbp) for moving averages, p-values (10^{-6}–10^{-2}) for statistical estimations and read coverage

(1–10×). We extracted optimal values of the parameters from the results of the simulation experiments. The performance of CNV_shape was then validated by experiments with five individual human genomes using the optimal parametric values. The assessment of the performance of CNV_shape was obtained by evaluating the false negative rate (FNR) and the false positive rate (FPR) on the basis of the Database of Genomic Variants (DGV; http://projects.tcag.ca/variation). The performance of CNV_shape was also compared to that of CNV-seq, which is one of the conventional CNV detection methods that use NGS data (Xie and Tammi, 2009). The results of the experiments showed that relatively small CNVs (> 1 kbp) can be precisely detected from read coverage data with very low coverage (> 1.7×). The results also presented that CNV_shape is very effective in reducing noises inherent in real data, as well as in detecting CNVs of various sizes (ranging from 1.0 kbp to 368 kbp) and types.

This paper consists of five sections. Section One describes the background, motivation and significance of the study. Section Two briefly reviews the CNV detection methods. Section Three introduces research methods, including the definition of a CNV based on variations in the shape of the read coverage data and the method of CNV_shape. Section Four presents the results of this study, including analysis of the experiments with simulated and real human data. Section Five discusses our important findings, highlights various research recommendations and limitations and makes suggestions for future studies.

2 Related work

There are two typical ways to detect CNVs: microarray-based methods (Agam et al., 2010; Fiegler et al., 2006; Ju et al., 2010; McCarroll et al., 2008; Redon et al., 2006) and sequence-based methods (Khaja et al., 2006; Medvedev et al., 2009; Snyder et al., 2010; Tuzun et al., 2005). Microarray-based methods are experimental methods that use an array such as an oligonucleotide array or a Bacterial Artificial Chromosome array. Recent CNV detection methods that use array comparative genetic hybridisation data are reviewed in Koike et al. (2011) and Lai et al. (2005). The methods determine CNVs by applying diverse techniques such as mixture models, hidden Markov models, maximum likelihood estimations, regression techniques, wavelet techniques and genetic algorithms. These array-based CNV detection methods are still cost-effective (Conrad et al., 2010), but are encountering the fundamental problem of cross-hybridisation.

There are two main approaches to sequence-based methods of CNV detection. One approach directly compares accurately completed sequence assemblies of genomes (Khaja et al., 2006; Tuzun et al., 2005). It can detect small and medium-sized CNVs because its resolution is higher than that of microarray-based methods; however, its application is not feasible, since there are few completed sequence assemblies of genomes due to the huge costs. The other approach uses NGS data. It detects CNVs by analysing read coverage data obtained by directly aligning reads onto a reference sequence (Abyzov et al., 2011; Alkan et al., 2009; Chiang et al., 2009; Ju et al., 2011; Medvedev et al., 2010; Mills et al., 2011; Xie and Tammi, 2009; Yoon et al., 2009). This approach needs short reads of high coverage to overcome the difficulty of detecting CNVs owing to the low signal-to-noise ratio that is characteristic of many platforms. CNV detection methods based on statistical modelling of a high coverage read data (20–30×) are suggested in Abyzov et al. (2011), Alkan et al. (2009) and Yoon et al. (2009). Several proposals include CNV detection methods based on

the relative difference between individual read coverage data: CNVs were assessed between a tumour sample and a normal sample in Chiang et al. (2009) and between two individuals, Dr. J. Craig Venter and Dr. James Watson, in Xie and Tammi (2009). The methods outlined in Chiang et al. (2009) and Xie and Tammi (2009) are applicable to short reads of relatively low coverage.

3 Method

This section describes the proposed method of detecting CNVs. Section 3.1 presents the problem definitions and resolutions and Section 3.2 explains the proposed method in detail.

3.1 Problem definitions

The notations and terms used to discuss the proposed CNV detection method are listed in Table 1 and the definitions below, respectively.

For comparison purposes, we use two different sequences in this study. One is a standard reference sequence: it is completely sequenced and has a well-known quality. The other is a test sequence. The test sequence is made of enormous sets of unassembled reads generated by NGS technology. Read coverage data is obtained by aligning the reads of the test sequence to the standard reference sequence. Then, it is analysed on the basis of shape variations detect the CNVs of the test sequence. Definition 1 describes a general definition of a CNV. Definitions 2 and 3 describe the terms used in the procedures to obtain the read coverage data. Definitions 4 through 9 present the terms used to detect CNVs in this study.

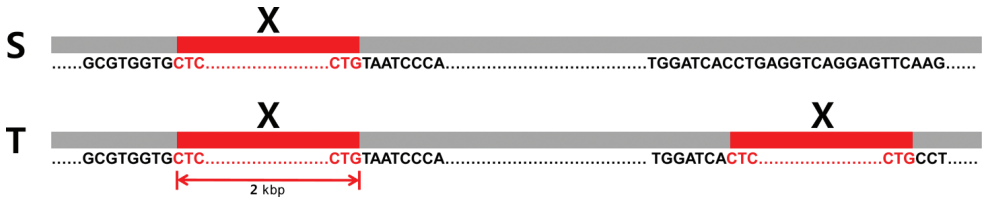
Definition 1 (CNV) *Given a reference sequence* $S = s_0 \ldots s_{n-1}$ *and a test sequence* $T = t_0 \ldots t_{m-1}$, *a subsequence* $X = x_0 \ldots x_{t-1}$ *of T is called a CNV if the copy number of X in T is different from that in S and the length t of the subsequence X is larger than or equal to a practical lower bound (≈1* kbp*) required for the length of a CNV.*

Figure 1 shows an example of a CNV as defined in Definition 1. Subsequence X is called a CNV because it is included in the sequences S and T, but occurs once in S and twice in T and further, it has length greater than 1 kbp.

Table 1 Summary of notations and terms

Symbol	*Short description*
$S_n = s_0 \ldots s_{n-1}$	Standard reference DNA sequence (n = size of S)
$X = x_0 \ldots x_{t-1}$	Subsequence of a sequence S (t = size of X, $1 \le t \le n$)
$T = t_0 \ldots t_{m-1}$	Test sequence (m = size of T)
$R_j = r_0 \ldots r_{j-1}$	Read of a test sequence T ($j = 1,\ldots, N$, N = total number of reads, d = size of R_j)
$C = c_0 \ldots c_{n-1}$	Read coverage data of a test sequence T obtained by aligning R_j ($j = 1,\ldots, N$) of T onto a reference sequence S
$mSH_{w,k}(C) = h_0 \ldots h_{s-1}$	Mean shift transform of C, s = total number of windows, w = window size, k = window shift size
$mSL_{w,k}(C) = l_0 \ldots l_{s-1}$	Mean slope transform of C, s = total number of windows, w = window size, k = window shift size

Figure 1 An example of a CNV region (see online version for colours)



Definition 2 (read alignment operator). *Let* $S = s_0 \ldots s_{n-1}$ *be a given standard reference sequence with length n,* $S_p = s_p \ldots s_{p+i-1}$ *be a subsequence of S with length l and* S'_p *be an extension of* S_p *by replacing r characters and or inserting g gaps. Let* $R_j = r_0^j \ldots r_{d-1}^i$ *be the* j^{th} *read sequence of N reads with length d of a given test sequence* $T = t_0 \ldots t_{m-1}$. *The definition of read alignment operator* $R_j = S_p$ *between the sequences* R_j *and* S_p *says that* R_j *is aligned to* S_p *if* S'_p *has the same sequence as* R_j *except for the gap positions and r and g are within tolerance limits.*

Definition 3 (read coverage sequence estimation). *Given a standard reference sequence* $S = s_0 \ldots s_{n-1}$, *subsequences* $\{S_p = s_p \ldots s_{p+l-1},\ 0 \le p,\ l < n\}$ *of S and the read sequences* $\{R_j = r_0^j \ldots r_{d-1}^j,\ 1 \le j \le N\}$ *of a given test sequence* $T = t_0 \ldots t_{m-1}$, *read coverage data* $C = c_0 \ldots c_{n-1}$ *of T is obtained by read coverage sequence estimation, which is defined as follows:*

$$c_i = \sum_{j=1}^{N} SC_j \text{ , where } SC_j = \begin{cases} 1, R_j \cong S_p\,(p \le i \le p+l-1) \\ 0, otherwise \end{cases}$$

Figure 2 shows an overview diagram of the processes of obtaining read coverage data by using the read alignment operator and the read coverage sequence estimation. The top part of Figure 2 shows the process of aligning the short reads $R_j = r_0^j \ldots r_{d-1}^j, j = 1, \ldots N$ to a reference sequence S by means of the read alignment operator ($R_j \cong S_p$). The bottom part of Figure 2 shows the process of obtaining the read coverage data $C = c_0 \ldots c_{n-1}$ from the aligned reads; the read coverage sequence estimation is used to count the number of reads aligning to each position of the reference sequence S.

Figure 2 An overview diagram of the processes of obtaining read coverage data (see online version for colours)

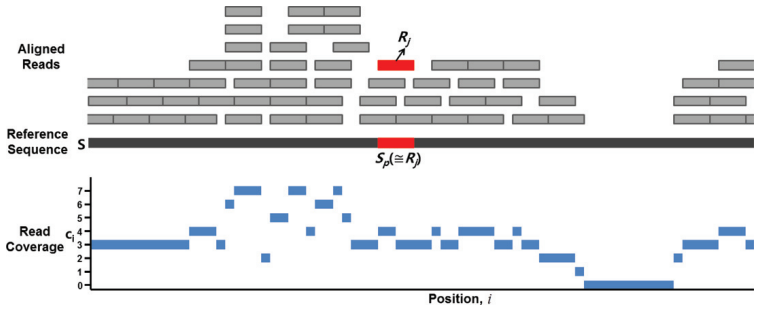
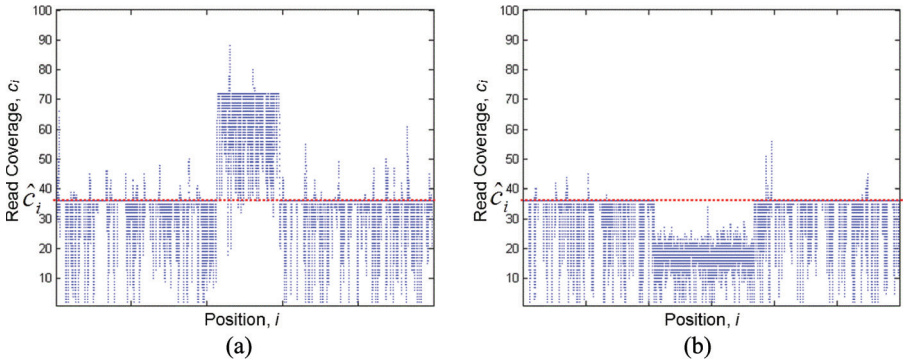


Figure 3 presents ideal cases of read coverage data for which noise effects are excluded. The *x*-axis represents position *i* along the reference sequence and the *y*-axis represents the values of the read coverage data c_i of the test sequence. The region which has relatively higher values than the overall average $\hat{c}_i$ of the read coverage data as shown in Figure 3(a) can be identified as a CNV gain since it evidently shows that the test sequence has more copies of the segment in the region than the reference sequence. The region which has relatively lower values than the overall average $\hat{c}_i$ of the read coverage data as shown in Figure 3(b) can be identified as a CNV loss since it evidently shows that the test sequence has less copies of the segment in the region than the reference sequence.

Figure 3 Examples of an ideal CNV: (a) gain region and (b) loss region (see online version for colours)



In general, CNVs are very hard to detect directly from the values of read coverage data in real human genome because real data have many zeros or unreliable high values in random positions due to sequencing errors, or because the coverage is not high enough. Therefore, we carried out two transforms, the mean shift transform and the mean slope transform to exclude the effects of noises or low coverage and extract CNVs based on the shape variations of the read coverage data. The mean shift transform and the mean slope transform are defined in Definitions 4 and 5, respectively.

Definition 4 (mean shift transform). *Let* $C = c_0 \ldots c_{n-1}$ *be the read coverage data of a test sequence* $T = t_0 \ldots t_{m-1}$ *over a given reference sequence* $S = s_0 \ldots s_{n-1}$. *The mean shift transform* $mSH_{w,k}(C)$ *of C is defined as follows:*

$$mSH_{w,k}(C) = h_0 \ldots h_{s-1}$$

where

$$h_i = \begin{cases} +1 & , \quad E(c_i) > c_{th}^{+} \\ 0 & , \quad c_{th}^{-} \le E(c_i) \le c_{th}^{+} \\ -1 & , \quad E(c_i) < c_{th}^{-} \end{cases}$$

Here, $E(c_i) = \frac{1}{w} \sum_{j=ik}^{ik+w-1} c_j$ *is the moving average of the ith window with size w and shift value k,* $s = \left\lceil \frac{n-w+1}{k} \right\rceil$ *is the total number of the windows and* c_{th}^{+} *and* c_{th}^{-} *are the upper and*

the lower bounds of the threshold value of $E(c_i)$, respectively. The values of c_{th}^{+} and c_{th}^{-} are calculated based on a statistical estimation with p-value p_h (< 0.05, two-tailed) derived from a central limit theorem.

Definition 5 (mean slope transform). *Let $C = c_0 \ldots c_{n-1}$ be the read coverage data of a test sequence $T = t_0 \ldots t_{m-1}$ over a given reference sequence $S = s_0 \ldots s_{n-1}$. The mean slope transform $mSL_{w,k}(C)$ of C is defined as follows:*

$$mSH_{w,k}(C) = l_0 \ldots l_{s-1}$$

where

$$l_i = \begin{cases} +1, \hat{c}_{th}^{-} \le dE_i \le \hat{c}_{th}^{+} \\ 0, otherwise \end{cases}$$

Here, $dE_i = \frac{1}{w'}\left(\sum_{j=ik+w-w'}^{ik+w-1} c_j - \sum_{j=ik}^{ik+w'-1} c_j \right)$ is a smoothed difference of the read coverage data in the ith window with size w and shift value k, w' is a given value in $0 < w' < w/2$, $s = \left\lceil \frac{n-w+1}{k} \right\rceil$ is the total number of the windows and $\hat{c}_{th}^{+}$ and $\hat{c}_{th}^{-}$ are the upper and the lower bounds of the threshold value of dE_i, respectively. The values of $\hat{c}_{th}^{+}$ and $\hat{c}_{th}^{-}$ are calculated based on a statistical estimation with p-value p_l (<0.01, two-tailed) derived from a central limit theorem.

In this method, the given read coverage data C are sampled via a window of size w and shift value k; in addition, the values of w and k are chosen for reducing the effects of the sequencing errors and low coverage. The mean shift transform $mSH_{w,k}(C)$ of C is then estimated to identify positions, called primary CNVs, with statistically significant differences in the values of the read coverage data. Next, the mean slope transform $mSL_{w,k}(C)$ extracts positions which are stationary, as described in Definition 5. Then, consecutive stationary positions with statistically significant differences in the values of the read coverage data can be identified as a candidate (CCNV) by using the results of the mean shift transform $mSH_{w,k}(C)$ and the mean slope transform $mSL_{w,k}(C)$. Definitions 6 and 7 describe the processes of obtaining CCNV gains and losses, respectively.

Definition 6 (CCNV gain). *Given the mean shift transform $mSH_{w,k}(C) = h_0 \ldots h_{s-1}$ and the mean slope transform $mSL_{w,k}(C) = l_0 \ldots l_{s-1}$ of the read coverage data $C = c_0 \ldots c_{n-1}$ of a given test sequence $T = t_0 \ldots t_{m-1}$, a CCNV gain $CCNV_i^{+}$ is defined as $CCNV_i^{+} = [p_j, p_j + m_j - 1]$ if the multiplication of h_i and l_i is non-zero positive for $i_{min} \le i \le i_{max}$, where $p_j = i_{min} \times k$, $p_j + m_j - 1 = i_{max} \times k + w - 1$ and $m_j \ge l_{cnv}$. Here, l_{cnv} is the minimum size for CNV identification and p_j and $p_j + m_j - 1$ are the starting and the ending positions of the j^{th} CCNV gain, respectively.*

Definition 7 (CCNV loss). *Given the mean shift transform $mSH_{w,k}(C) = h_0 \ldots h_{s-1}$ and the mean slope transform $mSL_{w,k}(C) = l_0 \ldots l_{s-1}$ of the read coverage data $C = c_0 \ldots c_{n-1}$ of a given test sequence $T = t_0 \ldots t_{m-1}$, a CCNV loss $CCNV_j^{-}$ is defined as $CCNV_j^{-} = [p_j, p_j + m_j - 1]$ if the multiplication of h_i and l_i is non-zero negative for $i_{min} \le i \le i_{max}$, where $p_j = i_{min} \times k$, $p_j + m_j - 1 = i_{max} \times k + w - 1$ and $m_j \ge l_{cnv}$. Here, l_{cnv} is the minimum size for CNV identification and p_j and $p_j + m_j - 1$ are the starting and the ending positions of the j^{th} CCNV loss, respectively.*

A CCNV satisfies the conditions of a CNV described in Definition 1. Now, a merging process, as the final step of the proposed method, is carried out for neighbouring CCNVs to finally determine the boundary of a putative CNV. The merging process is completed by an

evaluation of CCNV merging scores between neighbouring CCNVs and by CCNV merging operator as, defined in Definitions 8 and 9, respectively.

Definition 8 (CCNV merging score). *Given two neighbouring CCNV gains* $CCNV_j^+ = [p_j, p_j + m_j - 1]$ *and* $CCNV_{j+1}^+ = [p_{j+1}, p_{j+1} + m_{j+1} - 1]$*, the CCNV merging score* CMS_j^+ *is defined as follows:*

$$CMS_j^+ = \frac{\sum_{i=i_e}^{i_e} h_i}{p_{j+1} + m_{j+1} - p_j},$$

where $p_j = i_s \times k$ *and* $p_{j+1} + m_{j+1} - 1 = i_e \times k + w - 1$ *are the starting position of the jth CCNV gain and the ending position of the* $j + 1^{th}$ *CCNV gain, respectively. The merging score* CMS_j^- *of two neighbouring CCNV losses* $CMS_j^- = [p_j, p_j + m_j - 1]$ *and* $CMS_j^- = [p_j + 1, p_{j+1} + m_{j+1} - 1]$ *is defined in the same way as* CMS_j^+.

Definition 9 (CCNV merging operator). *Given two neighbouring CCNV gains* $CCNV_j^+$ *and* $CCNV_{j+1}^+$*, the CCNV merging operator : is defined by* $CCNV_j^+ : CCNV_{j+1}^+ = [p_j, p_{j+1} + m_{j+1} - 1]$*, if and only if the corresponding CCNV merging score* $CCNV_j^+$ *is larger than a lower bound* l_{sc}*; in addition, there are no other CCNV losses between the given two neighbouring CCNV gains. The term* $CCNV_j^- : CCNV_{j+1}^-$ *of the two neighbouring CCNV losses* $CCNV_j^-$ *and* $CCNV_{j+1}^-$ *is defined in the same way as the term* $CCNV_j^+ : CCNV_{j+1}^+$.

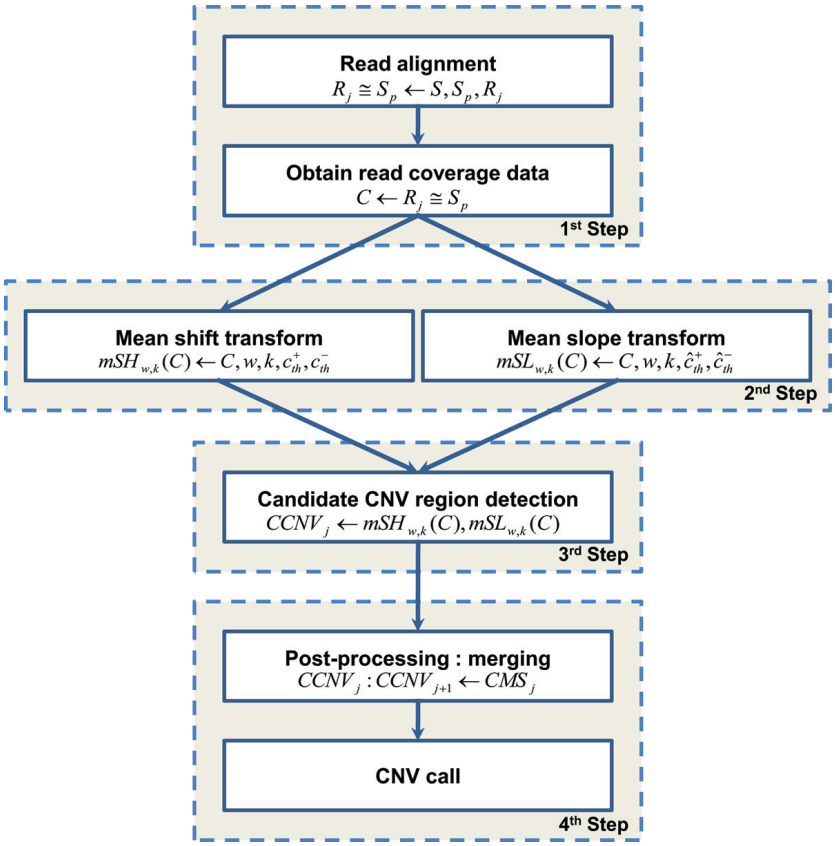
3.2 Proposed CNV detection method

Figure 4 shows a flowchart of the overall process of CNV_shape for detecting CNVs. The process has four steps. The first step is to obtain read coverage data $C = c_0 \ldots c_{n-1}$ of a test sequence T on a given reference sequence S by counting reads of the test sequence T aligned on each position of the reference sequence S. As in Definition 2, the read R_j is aligned onto a subsequence $S_p = s_p \ldots s_{p+l-1}$ of S if an extension S'_p of S_p has the same sequence as that of R_j except for g gap positions within tolerance limits. The read coverage data C of T is then obtained by the read coverage sequence estimation defined in Definition 3.

The second step is to get the mean shift transform $mSH_{w,k}(C) = h_0 \ldots h_{s-1}$ and the mean slope transform $mSL_{w,k}(C) = l_0 \ldots l_{s-1}$ of the read coverage data $C = c_0 \ldots c_{n-1}$ to exclude noise effects due to sequencing errors or effects of low coverage. The read coverage data $C = c_0 \ldots c_{n-1}$ is transformed into a sequence $\{h_i, i = 0, \ldots, s-1\}$ by means of the mean shift transform of Definition 4, where h_i has a value of ±1 or 0 depending on the average of the read coverage data $c_{ik} \ldots c_{ik+w-1}$ from that of the overall average of C. The fact that h_i has a value of +1 means that the average of the read coverage data $c_{ik} \ldots c_{ik+w-1}$ is significantly deviated in an upward direction and that a CNV gain may occur at the region $[ik, ik + w - 1]$ of S. Similarly, the fact that h_i has a value of –1 means that the average of the read coverage data $c_{ik} \ldots c_{ik+w-1}$ is significantly deviated in a downward direction and that a CNV loss may occur at the region $[ik, ik + w - 1]$ of S.

The read coverage data $C = c_0 \ldots c_{n-1}$ is also transformed into a sequence $\{l_i, i = 0, \ldots s - 1)$ by means of the mean slope transform of Definition 5. Here l_i has a value of +1 if the sequence $c_{ik} \ldots c_{ik+w-1}$ is stationary, or 0 if not, respectively. Now, the combination of $\{h_i, i = 0, \ldots, s - 1\}$ and $\{l_i, i = 0, \ldots, s - 1\}$ can be used to extract the regions of CNVs from the erroneous read coverage data $C = c_0 \ldots c_{n-1}$. The third step is to specify a CCNV gain CCNV_j^+ or a CCNV loss CCNV_j^- by using the combination of $\{h_i, i = 0, \ldots, s - 1\}$ and $\{l_i, i = 0, \ldots s - 1\}$, as in Definitions 6 and 7.

Figure 4 Flowchart of the CNV_shape processes of detecting CNVs (see online version for colours)



Finally, in the fourth step, a merging process is applied to all the CCNVs specified in the third step. The merging step of the post-processing is to reduce noise effects, such as fluctuations of read coverage depth, occurring when the size of the window is decreased in the processes of *mSH* and *mSL*. Originally one big CNV region may be detected as many separate small CNVs because of the fluctuations of read coverage depth. The merging of two consecutive candidate CNVs is proceeded if they have homogeneous features such as copy gains or losses and also if the distance between them is within a given specific length. The merging process can be iterative and optional. In the process, two neighbouring CCNV gains or two neighbouring CCNV losses are merged by using the merging scores of Definition 8 and the merging operator of Definition 9.

4 Results and discussion

Simulated data are used to estimate optimal values of input parameters in the CNV_shape method and real human data are used to evaluate the performance of CNV_shape. The performance evaluation was done by using FPR and FNR based on the CNV database of the

DGV and also by comparing the values of FPR and FNR of CNV_shape to those of CNV-seq, a conventional method.

4.1 Experimental method

A simulation data generator (SDG) was developed to generate simulated data. It generates a reference sequence and a test sequence, which contain CNVs of various sizes and locations; they also contain SNPs and short indels. The SDG inputs a given DNA sequence both as a reference sequence and as a test sequence. It then copies some of the CNVs of the sequence referring to the CNV database of the DGV and substitutes them in random positions of the reference sequence or the test sequence so that the test sequence has CNV gains or losses that differ in sizes and locations compared to the reference sequence. An indel is constructed by inserting or deleting a short sequence at a random position of the reference or the test sequence. For SNPs, the SDG replaces the nucleotides at random positions of the test sequence so that each of the replaced positions in the test sequence has a different nucleotide from that in the reference sequence and further, the sum of the replaced positions is about 3% of the total length of the test sequence. Once a reference sequence and a test sequence are generated, reads of the test sequence are generated by means of a paired-end method. Contig NT_077531.3 of human chromosome (chr.) 8 (NCBI Build 36.3) was used to generate 80 simulated sequences. Paired-end 36 bp reads were generated from each of the 80 simulated test sequences.

Chromosome 6 paired-end read data of 5 individuals, NA18507, NA18511, NA18570, NA18944 and NA10851, downloaded from the sites of the 1000 Genome project (http://www.1000genomes.org) and the TIARA database (http://tiara.gmi.ac.kr) were used for the experiments with real human data. The average coverage level of NA18511, NA18570, NA18507 and NA18944 were 1.7–2.3×. For NA10851, two different coverage levels, 5.6× and 25.01× were used.

The performance of CNV_shape was assessed by evaluating FPR and FNR on the level of nucleotide using the length of regions which overlap between CNVs of DGV database and the set of CNV regions from each algorithm of CNV_shape and CNV-seq. FPR is calculated by the ratio of regions incorrectly identified as CNVs to the whole non-CNV regions on the basis of the DGV database. FNR is calculated by the ratio of regions incorrectly identified as non-CNVs to the whole CNV region. The comprehensive information of CNV regions of human individuals disclosed through wet or dry lab experiments are reported in the CNV database of DGV in which region-specific and individual-specific searches are available. Even though the data submission in DGV is completed by filing and registering a given form on the web site without any procedures for verification, DGV is still recognised as one of the best reliable and the most frequently quoted references and is also incessantly updated by scientists around the world. All the information used in the validation was downloaded from DGV on March 25, 2010.

The CNV-seq (Xie and Tammi, 2009) detects CNVs on the basis of a statistical assessment of the read coverage data. It is applicable to data with relatively low level of coverage. However, CNV-seq has difficulty in classifying the types of detected CNVs because, as in micro-array methods, its method is based on the read coverage ratio of the test sequence to the control sequence. The performance of CNV_shape was compared with two types of implementation of CNV-seq, conventional CNV-seq and modified CNV-seq, where the modified CNV-seq was considered for having an input data set similar to CNV_shape,

that is, it has read coverage data of only a test sequence, but not a control sequence. The implementations of the conventional CNV-seq and the modified CNV-seq are explained below.

1 Conventional CNV-seq: The read coverage data of a test sequence and a control sequence were obtained by aligning reads of the test sequence and the control sequence to a reference sequence, respectively. The CNVs were then detected on the basis of a statistical assessment of the log ratio of the read coverage data of the test sequence to those of the control sequence. Here, we used the programs offered on the web site of (Xie and Tammi, 2009) (http://tiger.dbs.nus.edu.sg/cnv-seq).

2 Modified CNV-seq: The read coverage data of a test sequence were obtained by aligning the reads of the test sequence to a reference sequence. Next, the reads of the reference sequence were generated by applying a paired-end method to random positions of the reference sequence, where the coverage level of the reads of the reference sequence was about 36x. The read coverage data of a control sequence were then obtained by aligning the reads of the reference sequence to the reference sequence itself. Finally, CNVs were detected on the basis of a statistical assessment of the log ratio of the read coverage data of the test sequence to the read coverage data of the control sequence, as in the case in the conventional CNV-seq.

The Short Oligonucleotide Alignment Program (SOAP) (Li et al., 2008) was used for the alignment of the reads in the experiments with simulated and real human data. The alignment algorithm we used was a random match method which is one of various alignment algorithms that SOAP2 supports. The algorithm allowed at most 2 mismatches as a tolerable limit with regard to noise, such as sequence errors (Lee et al., 2009).

The experiments were conducted in the Windows 7 and CentOS 5.5 platforms, 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 CNV_shape was C# for .NET Framework 3.5.

4.2 Results and analysis

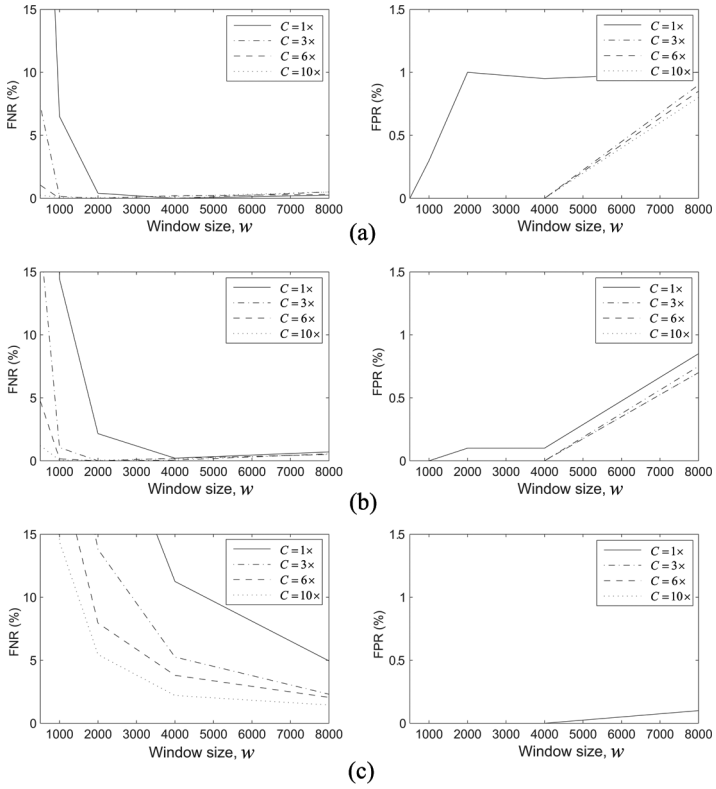
4.2.1 Experiment with simulated data

Figures 5, 6 and 7 show the results of experiments using simulated data with coverage levels, 1, 3, 6 and 10×, for various input parameters of CNV_shape, such as the window size w, the window shift size k and the p-values p_h and p_l. The p-values p_h and p_l are used for the statistical estimations of the mean shift transform and the mean slope transform, respectively. For each coverage level, the window sizes considered were 1, 2, 4, 6 and 8 kbp, the values of p_h were 0.01, 0.05 and 0.08 and the values of p_l were 0.000001, 0.0001 and 0.01. The window shift size k had three values, 1, $w/2$ and w for each window size w.

Figure 5 shows the FNR values (left) and the FPR values (right) of the simulated data as the window size increases for various coverage levels (1, 3, 6 and 10×) and for various p-values p_h (0.08, 0.05 and 0.01). Here, the window shift size k was given a value of 1 and the p-value p_l was given a value of 0.0001. Figures 5(a), 5(b) and 5(c) show the results for the p_h values of 0.08, 0.05 and 0.01, respectively. As shown in Figure 5, the FNR values decrease as the window size increases until the window size reaches a turning point w_{tp}^h and the FPR values are near zero up to the turning point w_{tp}^h, except in the case of 1× coverage. The decreasing rates of the FNR values become lower when the coverage or the p-value p_h is

reduced. There are slight increases in the values of the FNR and the FPR beyond the turning point w_{tp}^{h} . This increase occurs because the probability of detecting false CNVs as well as true CNVs increases as the window size increases.

Figure 5 Performance of CNV_shape for simulated data of coverage C in the range of 1× to 10× with different window sizes: here (a), (b) and (c) show the results for p_h-values of 0.08, 0.05 and 0.01, respectively (where p_h is the significance level for the mean shift transform). The window shift size k was set to 1 and the p-value p_l was set to 0.0001



Moreover, the probability of detecting small sizes of CNVs decreases owing to the greater smoothness effect as the window size increases. The turning point w_{tp}^{h} becomes higher when the coverage or the p-value p_h is lowered. The turning point w_{tp}^{h} was shown at the window size of 4,000, 2,000, 2,000 and 1,000 for coverage level 1, 3, 6 and 10×, respectively for the p-value ph of 0.05 in Figure 5(b). The optimal value of the window size w is now evaluated as the turning point w_{tp}^{h} for each coverage to have the best possible FNR and FPR values. Furthermore, the window size w should be less than or equal to 4,000 to avoid a sudden increase of the FPR for all the coverage levels, as shown in Figure 5(b). As a result, a window size of 4,000 at the p-value p_h of 0.05 is the optimal value of w for the best possible FNR and FPR values for all the coverage levels.

Figure 6 shows the FNR values (left) and the FPR values (right) of the simulated data as the window size w increases for various coverage levels (1, 3, 6 and 10×) and for various p-values p_l (0.01, 0.0001 and 10^{-6}). Here, the window shift size k was given a value of 1 and

the p-value p_h was given a value of 0.05. Figures 6(a), 6(b) and 6(c) show the results for the p-value p_l of 0.01, 0.0001 and 10^{-6}, respectively. As shown in Figure 6, the FNR values decrease as the window size increases until the window size reaches the turning point w_{tp}^l; in contrast, the FPR values are near zero up to the turning point w_{tp}^l, except in the case of 1× coverage. The decreasing rates of the FNR values become lower when the coverage is reduced or the p-value p_l is increased. There are slight increases in the values of the FNR and the FPR beyond the turning point w_{tp}^l. The turning point w_{tp}^l becomes higher when the coverage is lowered or the p-value p_l is increased. The turning point w_{tp}^l was shown at the window size of 4,000, 2,000, 2,000 and 1,000 for coverage levels 1, 3, 6 and 10x, respectively at the p-value of p_l, as shown in Figure 6(b). As a result, we chose a window size of 4,000 at the p-value p_l of 0.0001 as the optimal value of w for the best possible FNR and FPR values for all the coverage levels.

Figure 6 Performance of CNV_shape for simulated data of coverage C from 1× to 10× with different window sizes: here, (a), (b) and (c) show results for p_l-values of 0.01, 0.0001 and 0.000001 (where p_l is the significance level for the mean slope transform). The window shift size k was set to 1 and the p-value p_h was set to 0.05

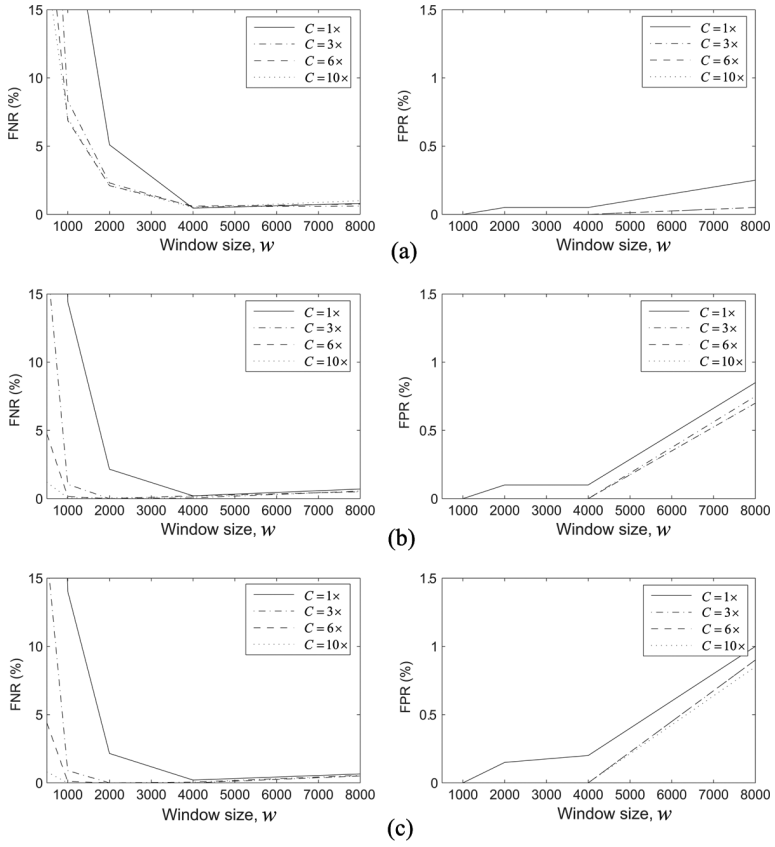
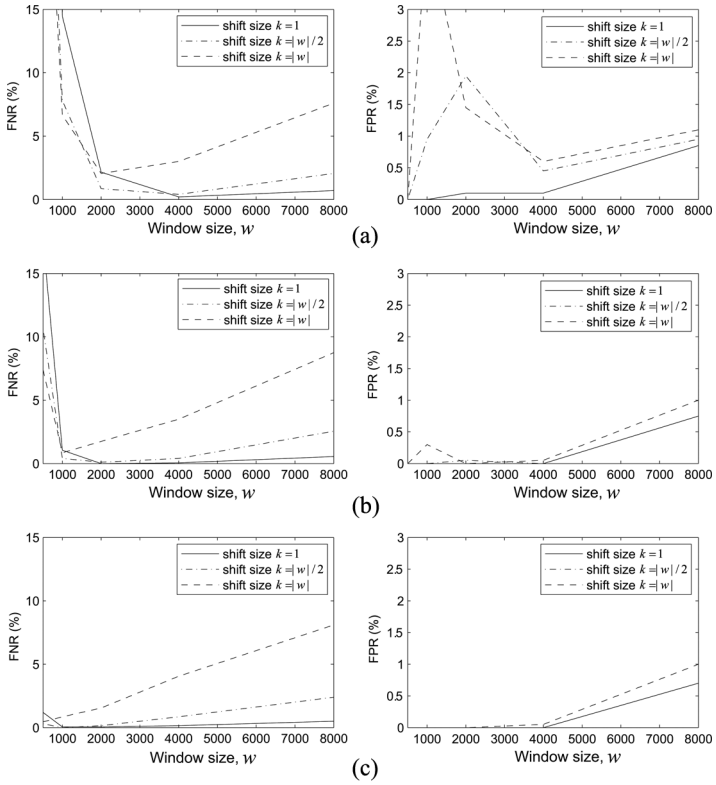


Figure 7 shows the FNR values (left) and the FPR values (right) as the window size w increases for various window shift sizes k (1, w/2 and w) and for various coverage levels (1, 3 and 10×). Here, the p-values p_h and p_l were given values of 0.05 and 0.0001, respectively.

Figures 7(a), 7(b) and 7(c) show the results for coverage levels 1, 3 and 10×, respectively. The results in Figure 7 show that the window shift size of $k = 1$ is the optimal value for all the coverage levels.

Figure 7 Performance of CNV_shape for various values of the window shift k with different window sizes: here, (a), (b) and (c) show the results for coverage levels of 1, 3 and 10×, respectively. The p-values p_h and p_l were set to 0.05 and 0.0001, respectively



In summary, the results of the experiments using the simulated data show that the best performance is attained at the window size of $w = 4{,}000$, the window shift size of $k = 1$ and the p-values of $p_h = 0.05$ and $p_l = 0.0001$, where the values of the FNR and the FPR, as a measure of the performance of the CNV_shape method, were in the range 0.05–0.2% and 0.001–0.1%, respectively.

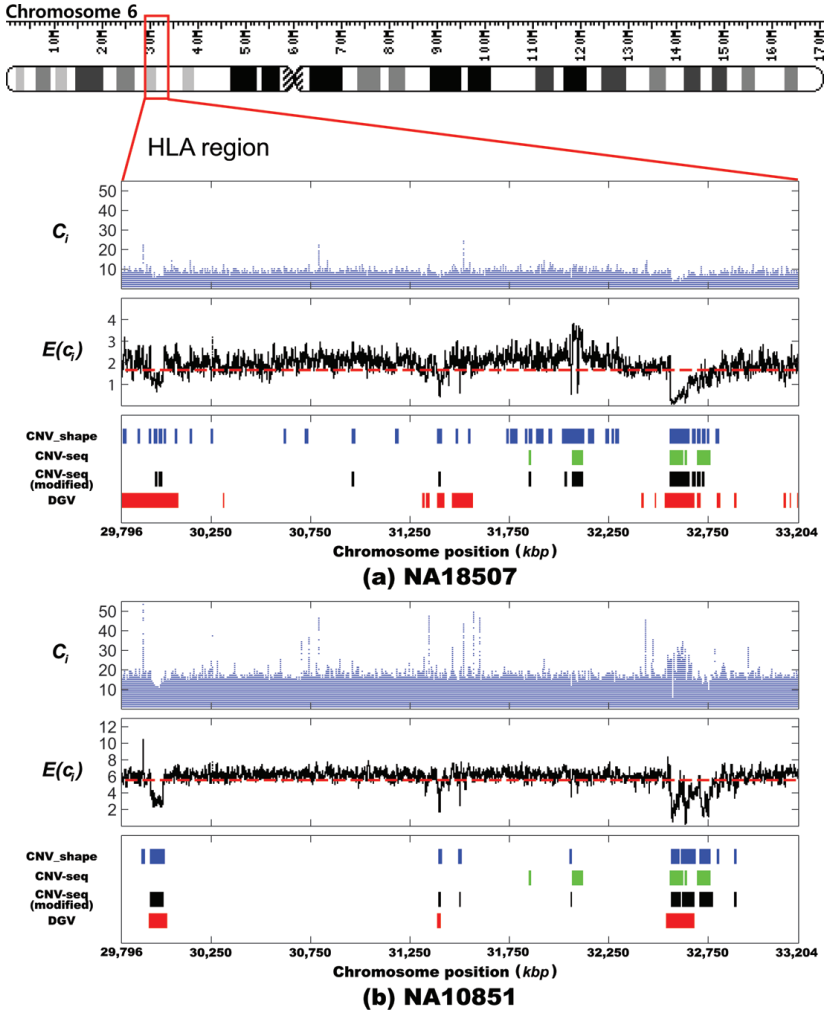
4.2.2 Experiment with hapmap samples

Sequencing data of five hapmap samples were used to assess the performance of CNV_shape; the optimal parametric values obtained from the experiments with simulated data were used as the values of the input parameters.

The assessment of the performance was accomplished by evaluating the FPR and the FNR on the basis of the CNV database of the DGV. The performance of CNV_shape was then compared with those of the conventional CNV-seq and the modified CNV-seq, as described in Section 4.1.

In the first experiment, we applied our method to read data from the human leukocyte antigen (HLA) region of chr. 6 of NA18507 and NA10851 to identify CNV regions. The average coverage level of the downloaded data was 1.7× for NA18507 and 5.6× for NA10851. The HLA region resides on the short arm of human chr. 6 and is 3.408 Mbp long; it contains around 200 genes related to the immune system function in humans. Figure 8 shows the experimental results of NA18507 (Fig. 8(a)) and NA10851 (Fig. 8(b)).

Figure 8 Example of CNV regions detected by the three methods in HLA regions of: (a) NA18507 and (b) NA10851. DGV represents the CNV regions reported in the CNV database of the DGV for HLA of chr. 6 (see online version for colours)



The top panels of Figures 8(a) and 8(b) show graphs of the read coverage data (Definition 3) of NA18507 and NA10851, respectively; the x axis is the position of chr. 6 and the y axis is the number of reads aligned to each position. The middle panels of Figures 8(a) and 8(b) display plots of the moving averages of the read coverage data of NA18507 and

NA10851, respectively. The moving averages are obtained during the mean shift transform with a window size w of 4,000 and a window shift size k of 1. As noted with respect to the top and middle panels, the graphs of the middle panel show variations in the shape of the read coverage data more clearly than those of the top panels. The distribution of the graph in the middle panel of Figure 8(b) is more stabilised than that of Figure 8(a), reflecting the difference in the coverage levels of NA18507 (1.7×) and NA10851 (5.6×).

The bottom panels of Figures 8(a) and 8(b) show the CNVs detected on NA18507 and NA10851, respectively, by CNV_shape; they also show the CNVs detected by both the conventional CNV-seq and the modified CNV-seq. Here, we used reads of NA10851 and NA18507 for the test and the control sequences, respectively, in the conventional CNV-seq. The CNVs reported in the CNV database of the DGV are also shown in the bottom panels. As plotted in the bottom panels of Figures 8(a) and 8(b), a total of 606,194 bp and 236,332 bp of the CNVs are reported in the CNV database of the DGV for HLA of chr. 6 on NA18507 and NA10851, respectively. As observed in the middle and bottom panels of Figure 8, the regions with relatively high (low) values of read coverage data in the graphs of the middle panels are mapped to the CNVs detected by CNV_shape in the bottom panels, confirming the intuitive and reliable nature of the CNV_shape method, which is based on variations in the shape of the read coverage data. CNV_shape accurately detects the CNV gains and losses for both NA18507 and NA10851. Furthermore, as shown in the plot (Fig. 8(a)) of CNVs detected by CNV_shape, many small regions are detected as CNVs on NA18507 with a relatively low coverage (1.7×). On the other hand, the CNVs detected by the conventional CNV-seq on NA18507 and NA10851 fail to include regions in which shape variations in the read coverage data of the test sequence are the same as those of the control sequence. Neither the conventional nor the modified CNV-seq can verify the types of detected CNVs because, as in micro-array methods, the detection of the CNV gains and losses is based on the ratio of the coverage of the test sequence to that of the control sequence. For example, as shown in the bottom panels of Figures 8(a) and 8(b), CNV-seq failed to detect a region of NA18507 (chr. 6, 31,393,061 bp to 31,423,019 bp) and a region of NA10851 (chr. 6, 31,394,254 bp to 31,404,429 bp), which are both reported as CNV regions in the CNV database of the DGV. Furthermore, the region detected by CNV-seq (namely, chr. 6, 32,712,915 bp to 32,759,142 bp) is estimated to be a CNV gain on NA18507 and at the same time, a CNV loss on NA10851, even though the same region can be verified as a CNV loss from the graphs of the middle panels of Figures 8(a) and 8(b). FNR (FPR) values of 62.905% (11.47%) and 26.40% (2.73%) were derived for NA18507 and NA10851 in CNV_shape on the basis of the CNV database of the DGV. In contrast, the conventional CNV-seq yielded FNR (FPR) values of 87.37% (3.63%) and 71.35% (3.48%) and the modified CNV seq yielded FNR (FPR) values of 77.075% (3.07%) and 39.01% (1.78%) for NA18507 and NA10851, respectively.

In the second experiment, we applied our method to read data from the whole region of human chr. 6 (170.899 Mbp long) of NA18507 and NA10851. For chr. 6 of NA18507 and NA10851, 354 (min, max and total sizes are 1,001 bp, 368,592 bp and 2,464,870 bp, respectively) and 34 (min, max and total sizes are 1,018 bp, 142,220 bp and 852,185 bp, respectively) CNVs are reported on the DGV database, respectively.

Table 2 describes the comparative performance results of CNV_shape, the conventional CNV-seq and the modified CNV-seq. Here, columns, min and max, represent the smallest and the largest sizes respectively of CNVs correctly detected on the basis of DGV by each method. The total fraction of detected CNV regions overlapping with those of DGV is also given in parentheses. Columns, gain and loss, represent the total sum of the lengths of CNV

gains and losses, respectively by each method. FNR values of 56.89% and 24.94% were derived for NA18507 and NA10851 in CNV_shape on the basis of the CNV database of the DGV. On the other hand, CNV-seq yielded FNR values of 94.385% and 72.82% and the modified CNV-seq yielded FNR values of 73.275% and 40.00% for NA18507 and NA10851, respectively.

As given in Table 2, the sizes of CNVs detected by CNV_shape on NA18507 and NA10851 are in the range of 1 kbp to 368 kbp and 1 kbp to 142 kbp, respectively. Figure 9 shows the numbers of CNVs detected by the CNV_shape algorithm on chromosome 6 of NA18507 (Fig. 9(a)) and NA10851 (Fig. 9(a)) over the CNV size. Here, the boxes coloured red show the numbers of CNVs reported in DGV and the boxes coloured blue show the numbers of CNVs detected by the CNV_shape algorithm and overlapped more than 50% with those of DGV. For chromosome 6 of NA18507, 354 CNVs are reported on the DGV database, 74 of which overlap more than 50% with CNV regions called by the CNV_shape algorithm. For chromosome 6 of NA10851, 34 CNVs are reported on the DGV database, 21 of which overlap more than 50%, with CNV regions called by the CNV_shape algorithm. In particular, 85% of CNV regions reported for chromosome 6 of NA18507 on DGV belong to sizes less than ~ 5 kbp. From the comparative analysis, we found that CNVs with size less than ~ 5 kbp are hard to detect for chromosome 6 of NA18507 due to their extremely low coverage ~ 1.7× and their CNV size distribution. Therefore, we can suggest that the CNV_shape algorithm may find small CNVs fairly well if the coverage of a sample is not very low.

Figure 9 Distribution of CNV sizes of: (a) NA18507 and (b) NA10851. Here, the boxes coloured red show the numbers of CNVs reported in DGV and the boxes coloured blue show the numbers of CNVs detected by the CNV_shape algorithm and overlapping more than 50% with those of DGV (see online version for colours)

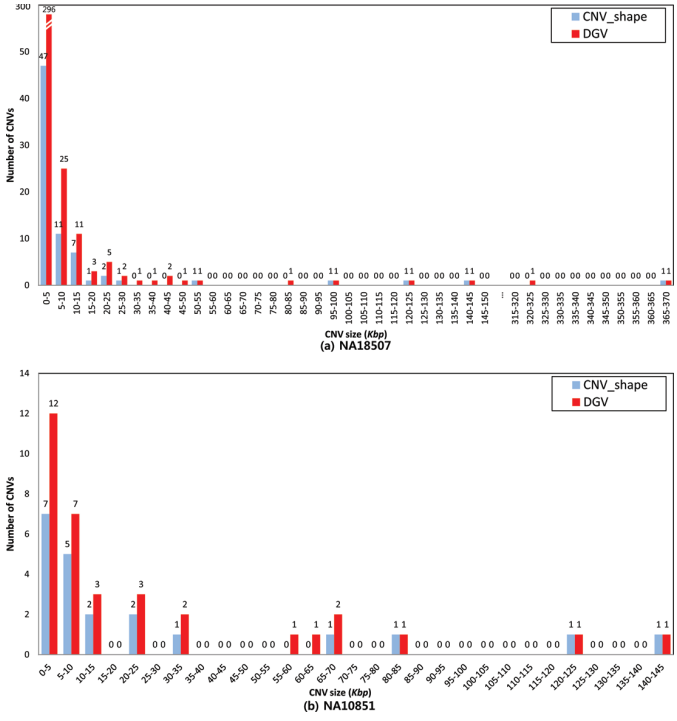


Table 2 Comparative summary of performance in the whole regions of human chromosome 6 of NA18507 and NA10851

Method	NA18507						NA10851					
	Detected region (kbp)				FPR (%)	FNR (%)	Detected region (kbp)				FPR (%)	FNR (%)
	min	max	gain	loss			min	max	gain	loss		
CNV_shape	1.0 (100%)	368 (96.4%)	8894	4458	7.48	56.89	1.0 (100%)	142 (92.2%)	1020	931	0.92	24.94
CNV-seq	1.6 (72.1%)	143 (47.0%)	128	328	0.18	94.38	6.2 (100%)	142 (7.2%)	63	288	0.17	72.82
Modified CNV-seq	1.2 (30.4%)	368 (69.4%)	823	1833	1.18	73.27	1.3 (100%)	142 (55.7%)	538	757	0.54	40.00

Table 3 Performance comparison of three methods on chromosome 6 sequence data of five individuals

	NA18507 (1.7X)		*NA18511 (1.8X)*		*NA18570 (2.2 X)*		*NA18944 (2.3X)*		*NA10851*			
									(5.6X)		*(25.01X)*	
Method	*FPR*	*FNR*	*FPR*	*FNR*	*FPR*	*FNR*	*FPR*	*FNR*	*FPR*	*FNR*	*FPR*	*FNR*
CNV_shape	7.48	56.89	5.14	46.19	3.11	48.81	5.55	42.53	0.92	24.94	5.62	9.19
CNV-seq	0.18	94.38	0.13	86.29	0.11	81.99	0.10	82.78	0.17	72.82	2.02	75.96
Modified CNV-seq	1.15	73.27	0.56	57.68	0.52	54.53	0.78	46.32	0.54	40.00	5.74	52.17

These results confirm that CNV_shape is superior to the conventional CNV-seq and the modified CNV-seq in terms of detecting CNVs of various sizes. We can deduce, therefore, that CNV_shape is very effective at reducing the noise inherent in the read coverage data and in detecting CNVs of various sizes and types.

In the third experiment, we applied additionally our method on chr. 6 data of three individuals at low coverage. For comparison, we also used a high coverage read data of NA10851 generated in the paper (Park et al., 2010). Table 3 summarises the results of the comparative analysis. As shown in the table, the overall FNR and FPR of CNV_shape are in the range of 24.94–56.89% and 0.92–7.48%, respectively for relatively low coverage data (1.7–5.6×). Table 3 also shows that the values of FNR and FPR of CNV_shape are 9.19% and 5.62%, respectively for NA10851 with very high coverage level (25.01×). The results suggest that FNR and FPR values may be decreased as the coverage level of samples increases. Furthermore, CNV_shape has relatively good values of FPR and FNR at relatively low coverage levels (1.7–5.6×), as well as at a high coverage level (25.01×). CNV_shape outperforms the other methods by as much as 8.18–87.90%.

5 Conclusion

We have proposed a novel CNV detection method, CNV_shape, which is based on variations in the shape of read coverage data obtained by aligning NGS data onto a reference sequence. The proposed method carries out two transforms, the mean shift transform and the mean slope transform, to extract patterns of the shape variations of the read coverage data in the detection of CNVs. The mean shift and the mean slope transforms eliminated problems associated with the occurrence of many zeros or unreliably high-valued coverage values in random positions due to sequencing errors or low coverage. The performance of CNV_shape was assessed through experiments with simulated data and real human data i and the performance results were compared with those of existing CNV detection methods, namely conventional CNV-seq and the modified CNV-seq. The results show that the CNV_shape method, which is based on shape variations in the read coverage data, effectively detects CNV gains and losses on a relatively low coverage data. The CNV_shape also outperforms the conventional CNV-seq and the modified CNV-seq in terms of detecting CNVs with a wider range of sizes. The proposed method uses only a test sequence and does not need another sequence as a control. Therefore, the proposed method could be very useful for detecting homogeneous and heterogeneous CNVs in many individual sets of data. In other words, the detection of homogeneous and heterogeneous CNVs might be possible only by comparing the patterns of shape variations derived when the mean shift transform and the mean slope transform are applied to the read coverage data of many individual sets of data. In an effort to improve the performance of CNV_shape, we are currently investigating a method for the detection of homogeneous and heterogeneous CNVs from many individual sets of data.

Acknowledgements

This research 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).

References

Abyzov, A., Urban, A.E., Snyder, M. and Gerstein, M. (2011) 'CNVnator: an approach to discover, genotype and characterize typical and atypical CNVs from family and population genome sequencing', *Genome Research*, Vol. 21, No. 6, pp.974–984.

Agam, A., Yalcin, B., Bhomra, A., Cubin, M., Webber, C., Holmes, C., Flint, J. and Mott, R. (2010) 'Elusive Copy Number Variation in the mouse genome', *PLoS ONE*, Vol. 5, No. 9, p.e12839.

Alkan, C., Kidd, J.M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., Kitzman, J.O., Baker, C., Malig, M., Mutlu, O., Sahinalp, S.C., Gibbs, R.A. and Eichler, E.E. (2009) 'Personalized copy number and segmental duplication maps using next-generation sequencing', *Nature Genetics*, Vol. 41, No. 10, pp.1061–1067.

Antonina, M. and Bhubaneswar, B.M. (2010) 'On a novel coalescent model for genome-wide evolution of Copy Number Variations', *International Journal of Data Mining and Bioinformatics*, Vol. 4, No. 3, pp.300–315.

Chiang, D.Y., Getz, G., Jaffe, D.B., O'Kelly, M.J., Zhao, X., Carter, S.L., Russ, C., Nusbaum, C., Meyerson, M. and Lander, E.S. (2009) 'High-resolution mapping of copy-number alterations with massively parallel sequencing', *Nature methods*, Vol. 6, No. 1, pp.99–103.

Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C.H., Kristiansson, K., Macarthur, D.G., Macdonald, J,R., Onyiah, I., Pang, A.W., Robson, S., Stirrups, K., Valsesia, A., Walter, K., Wei, J., Wellcome Trust Case Control Consortium, Tyler-Smith, C., Carter, N.P., Lee, C., Scherer, S.W. and Hurles, M.E. (2010) 'Origins and functional impact of Copy Number Variation in the human genome', *Nature*, Vol. 464, No. 7289, pp.704–712.

Fiegler, H., Redon, R., Andrews, D., Scott, C., Andrews, R., Carder, C., Clark, R., Dovey, O., Ellis, P., Feuk, L., French, L., Hunt, P., Kalaitzopoulos, D., Larkin, J., Montgomery, L., Perry, G.H., Plumb, B.W., Porter, K., Rigby, R.E., Rigler, D., Valsesia, A., Langford, C., Humphray, S.J., Scherer, S.W., Lee, C., Hurles, M.E. and Carter, N.P. (2006) 'Accurate and reliable high-throughput detection of Copy Number Variation in the human genome', *Genome Research*, Vol. 16, No. 12, pp.1566–1574.

Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M., Donahoe, P., Qi, Y., Scherer, S. and Lee, C. (2004) 'Detection of large-scale variation in the human genome', *Nature Genetics*, Vol. 36, No. 9, pp.949–951.

International HapMap (2003) 'The international hapmap project', *Nature*, Vol. 426, No. 6968, pp.789–796.

Ju, Y.S., Hong, D., Kim, S., Park, S.S., Kim, S., Lee, S., Park, H., Kim, J.I. and Seo, J.S. (2010) 'Reference-unbiased copy number variant analysis using CGH microarrays', *Nucleic Acids Research*, Vol. 38, No. 20, p.e190.

Ju, Y.S., Kim, J.I., Kim, S., Hong, D., Park, H., Shin, J.Y., Lee, S., Lee, W.C., Kim, S., Yu, S.B., Park, S.S., Seo, S.H., Yun, J.Y., Kim, H.J., Lee, D.S., Yavartanoo, M., Kang, H.P., Gokcumen, O., Govindaraju, D.R., Jung, J.H., Chong, H., Yang, K.S., Kim, H., Lee, C. and Seo, J.S. (2011) 'Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals', *Nature Genetics*, Vol. 43, No. 8, pp.745–752.

Jun, D., Yuzhen, G., Yan, H., Yifeng, Z., Moli, H. and Haiyan, L. (2011) 'Screening SNPs residing in the micro RNA-binding sites of Hepatocellular Carcinoma related genes', *International Journal of Data Mining and Bioinformatics*, Vol. 5, No. 1, pp.1–21.

Khaja, R., Zhang, J., MacDonald, J.R., He, Y., Joseph-George, A.M., Wei, J., Rafiq, M.A., Qian, C., Shago, M., Pantano, L., Aburatani, H., Jones, K., Redon, R., Hurles, M., Armengol, L., Estivill, X., Mural, R.J., Lee, C., Scherer, S.W. and Feuk, L. (2006) 'Genome assembly comparison identifies structural variants in the human genome', *Nature Genetics*, Vol. 38, No. 12, pp.1413–1418.

Koike, A., Nishida, N., Yamashita, D. and Tokunaga, K. (2011) 'Comparative analysis of Copy Number Variation detection methods and database construction', *BMC Genetics*, Vol. 12, No. 29.

Lai, W.R., Johnson, M.D., Kucherlapati, R. and Park, P.J. (2005) 'Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data', *Bioinformatics*, Vol. 21, No. 19, pp.3763–3770.

Lee, K., Yoon, J., Hong, D., Hong, S. and Seong, D. (2009) 'An efficient read alignment method to detect genetic structural variations', *Paper Presented at the Internet and Multimedia Systems and Applications*, 13–15 July 2009, Cambridge, United Kingdom.

Li, R., Li, Y., Kristiansen, K. and Wang, J. (2008) 'SOAP: Short Oligonucleotide Alignment Program', *Bioinformatics*, Vol. 24, No. 5, pp.713–714.

McCarroll, S.A., Kuruvilla, F.G., Korn, J.M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M.H., de Bakker, P.I., Maller, J.B., Kirby, A., Elliott, A.L., Parkin, M., Hubbell, E., Webster, T., Mei, R., Veitch, J., Collins, P.J., Handsaker, R., Lincoln, S., Nizzari, M., Blume, J., Jones, K.W., Rava, R., Daly, M.J., Gabriel, S.B. and Altshuler, D. (2008) 'Integrated detection and population-genetic analysis of SNPs and Copy Number Variation', *Nature Genetics*, Vol. 40, No. 10, pp.1166–1174.

Medvedev, P., Fiume, M., Dzamba, M., Smith, T. and Brudno, M. (2010) 'Detecting Copy Number Variation with mated short reads', *Genome Research*, Vol. 20, No. 11, pp.1613–1622.

Medvedev, P., Stanciu, M. and Brudno, M. (2009) 'Computational methods for discovering structural variation with next-generation sequencing', *Nature methods*, Vol. 6, No. 11, pp.S13–S20.

Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., Abyzov, A., Yoon, S.C., Ye, K., Cheetham, R.K., Chinwalla, A., Conrad, D.F., Fu, Y., Grubert, F., Hajirasouliha, I., Hormozdiari, F., Iakoucheva, L.M., Iqbal, Z., Kang, S., Kidd, J.M., Konkel, M.K., Korn, J., Khurana, E., Kural, D., Lam, H.Y., Leng, J., Li, R., Li, Y., Lin, C.Y., Luo, R., Mu, X.J., Nemesh, J., Peckham, H.E., Rausch, T., Scally, A., Shi, X., Stromberg, M.P., Stutz, A.M., Urban, A.E., Walker, J.A., Wu, J., Zhang, Y., Zhang, Z.D., Batzer, M.A., Ding, L., Marth, G.T., McVean, G., Sebat, J., Snyder, M., Wang, J., Ye, K., Eichler, E.E., Gerstein, M.B., Hurles, M.E., Lee, C., McCarroll, S.A. and Korbel, J.O. (2011) 'Mapping Copy Number Variation by population-scale genome sequencing', *Nature*, Vol. 470, No. 7332, pp.59–65.

Park, H., Kim, J-I., Ju, Y-S., Gokcumen, O., Mills, R.E., Kim, S., Lee, S., Suh, D., Hong, D., Kang, H.P., Yoo, Y.J., Shin, J-Y., Kim, H-J., Yavartanoo, M., Chang, Y.W., Ha, J-S., Chong, W., Hwang, G-R., Darvishi, K., Kim, H., Yang, S.J., Yang, K-S., Kim, H., Hurles, M.E., Scherer, S.W., Carter, N.P., Tyler-Smith, C., Lee, C. and Seo, J-S. (2010) 'Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing', *Nature Genetics*, Vol. 42, No. 5, pp.400–405.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., Cho, E.K., Dallaire, S., Freeman, J.L., Gonzalez, J.R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J.R., Marshall, C.R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M.J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D.F., Estivill, X., Tyler-Smith, C., Carter, N.P., Aburatani, H., Lee, C., Jones, K.W., Scherer, S.W. and Hurles, M.E. (2006) 'Global variation in copy number in the human genome', *Nature*, Vol. 444, No. 7118, pp.444–454.

Snyder, M., Du, J. and Gerstein, M. (2010) 'Personal genome sequencing: current approaches and challenges', *Genes & Development*, Vol. 24, No. 5, pp.423–431.

Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., Olson, M.V. and Eichler, E.E. (2005) 'Fine-scale structural variation of the human genome', *Nature Genetics*, Vol. 37, No. 7, pp.727–732.

Xie, C. and Tammi, M.T. (2009) 'CNV-seq, a new method to detect Copy Number Variation using high-throughput sequencing', *BMC Bioinformatics*, Vol. 10, No. 1, p.80.

Yoon, S., Xuan, Z., Makarov, V., Ye, K. and Sebat, J. (2009) 'Sensitive and accurate detection of copy number variants using read depth of coverage', *Genome Research*, Vol. 19, No. 9, pp.1586–1592.