Computers in Biology and Medicine 42 (2012) 468–473

Contents lists available at SciVerse ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

CNV detection method optimized for high-resolution arrayCGH by normality test

Jaegyoon Ahn [a], Youngmi Yoon [b], Chihyun Park [a], Sanghyun Park [a,*]

[a] Department of Computer Science, Yonsei University, 134 Sinchon-dong, Seodaemun-gu, Seoul 120-749, South Korea
[b] Department of Computer Engineering, Gachon University, Gyeonggi-do, South Korea

ARTICLE INFO

Article history:
Received 15 December 2010
Accepted 27 December 2011

Keywords:
Data mining
Copy number variation
High-resolution arrayCGH
Genome analysis
Normality test

ABSTRACT

High-resolution arrayCGH platform makes it possible to detect small gains and losses which previously could not be measured. However, current CNV detection tools fitted to early low-resolution data are not applicable to larger high-resolution data. When CNV detection tools are applied to high-resolution data, they suffer from high false-positives, which increases validation cost. Existing CNV detection tools also require optimal parameter values. In most cases, obtaining these values is a difficult task.

This study developed a CNV detection algorithm that is optimized for high-resolution arrayCGH data. This tool operates up to 1500 times faster than existing tools on a high-resolution arrayCGH of whole human chromosomes which has 42 million probes whose average length is 50 bases, while preserving false positive/negative rates. The algorithm also uses a normality test, thereby removing the need for optimal parameters. To our knowledge, this is the first formulation for CNV detecting problems that results in a near-linear empirical overall complexity for real high-resolution data.

1. Introduction

CNV has been a hot research topic for the last several years, and recent studies have shown that CNV accounts for a significant proportion of normal phenotypic variation, as well as variation resulting in disease susceptibility [1–4]. Understanding the genetic basis of phenotypic variation and disease susceptibility in humans is an important goal in genetics, and the accurate and thorough detection of CNVs is also important. Most CNVs found using arrayCGH or SNParray are typically greater than 100 kb in length. However it is likely that CNVs shorter than 100 kb are much more numerous than currently believed [5].

The best solution for accurately detecting shorter CNVs is to compare two or more whole human genomes; however, sequencing whole genomes is often cost-prohibitive. A more practical solution is to increase the microarray resolution by shortening the length of each arrayCGH probes, or by reducing the inter-probe distance of SNParrays. CNVs are defined as gains or losses that are greater than 1 kb of a genomic DNA sequence [6,7]. This means that, theoretically, all CNVs could be detected, as long as the length of each probe of the microarray is less than 1 kb.

The Sanger Institute (http://www.sanger.ac.uk) has published high-resolution arrayCGH data. They possess 42 million probes spread across 2.1 million probe arrays with an average probe length of 50 bases. Unfortunately, most existing CNV detection algorithms cannot be applied to this high-resolution arrayCGH data without modification primarily because the millions of probes associated with high-resolution arrayCGH data result in a prohibitively long algorithmic run time.

For low resolution data, even one or two probes can constitute one CNV. The existing CNV detection algorithms developed mostly for low resolution data do not limit the number of probes per CNV. However, in the case of high resolution, one or two probes are too short to compose one CNV.

Existing CNV detection algorithms are also problematic because they are sensitive to user parameters. Because CNVs have not been precisely validated, those parameters cannot be optimized through cross validation. Accordingly, insensitivity to the parameters, a wide range of optimal parameters, or a limited number of parameters are desirable in a CNV detection algorithm.

This study proposes a novel CNV detection algorithm that divides probes into intervals and uses a normality test to determine whether those intervals are included in CNVs or not. This algorithm (1) applies to arrayCGH of any resolution by considering the size of probe, (2) has a reasonable runtime even for very high-resolution data, (3) requires only one parameter to specify the minimal length of the desired CNV, and (4) has a low false negative rate as discussed in the Results section.

2. Methods

2.1. Algorithm overview

An overview of the CNV detection algorithm proposed in this study is shown in Fig. 1. First, the whole probes are normalized

* Corresponding author. Tel.: +82 2 2123 5714; fax: +82 2 365 2579.
E-mail address: sanghyun@cs.yonsei.ac.kr (S. Park).

0010-4825/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2011.12.015

and clustered into intervals. Details of this preprocessing are described in Section 2.2. Then, a normality test is used to determine whether each interval is a CNV. Section 2.3 describes the details of the normality test. Finally, CNVs are constructed using candidate CNVs obtained by the normality test, as described in Section 2.4.

2.2. Preprocessing

Normalization of the whole probes within one chromosome is achieved via Z-transform [8]. After normalizing, the whole probes within each chromosome are divided into several intervals. For example, when the average probe size is 50 bases and the interval size is 1000 bases, the intervals are composed of approximately 20 continuous probes. The log ratio value of the interval refers to the average of the log ratio values of the probes that form the interval.

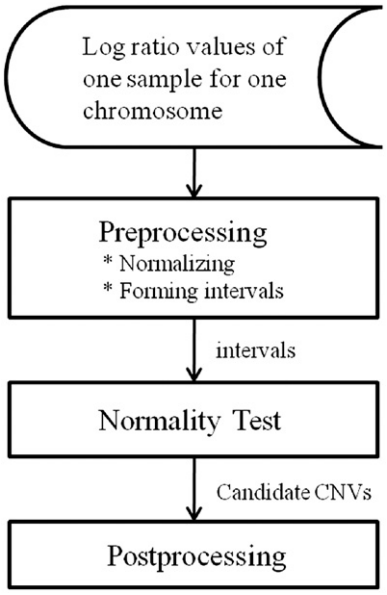

Fig. 1. Overview of the CNV detection algorithm.

If the resolution of arrayCGH data is high (i.e., the average size of the probe is less than 1000 bases), then the number of probes in one chromosome is too large to process. In that case, one probe cannot form a CNV assuming that the size of a CNV is greater than or equal to 1000 bases. Therefore, it is desirable to divide whole probes within one chromosome into intervals of probes. However, in the case of lower resolution arrayCGH data with a probe size greater than 1000 bases, each probe is considered one interval.

Interval size, which is the only parameter of the algorithm proposed in this study, can be thought of the minimal length of the desired CNV in bases we want to find. If not specified, the default value of the interval size is 1000 bases. If the resolution of an input file is greater than 1 kb, the interval size should be adjusted. For example, if the average length of a probe is 10 kb, then the minimum length of a CNV that can be detected would be <10 kb. Therefore, the interval size should be set to 10,000 bases or greater. If the size of interval is reduced, CNVs that are composed of one or two probes (50–100 bases on average) can be found using the method described in this study, although there is a high likelihood that these shorter CNVs will be false positives.

2.3. Normality test

We can assume that the log ratio values of arrayCGH data whose reference and test samples are same would follow normal distribution with an average of zero [9], due to random noise during WGTP experiment. In the case of arrayCGH data whose reference and test samples are not same, the average of the log ratio values of the probes that are included in the structural variation (including CNVs) between reference and test would not be equal to zero. So we can think that the log ratio values of these regions disturb the normal distribution of the total log ratio values. As the level of variation increases, the degree of normal distribution decreases. Therefore, it seems safe to assume that, as log ratio values suspected to be due to

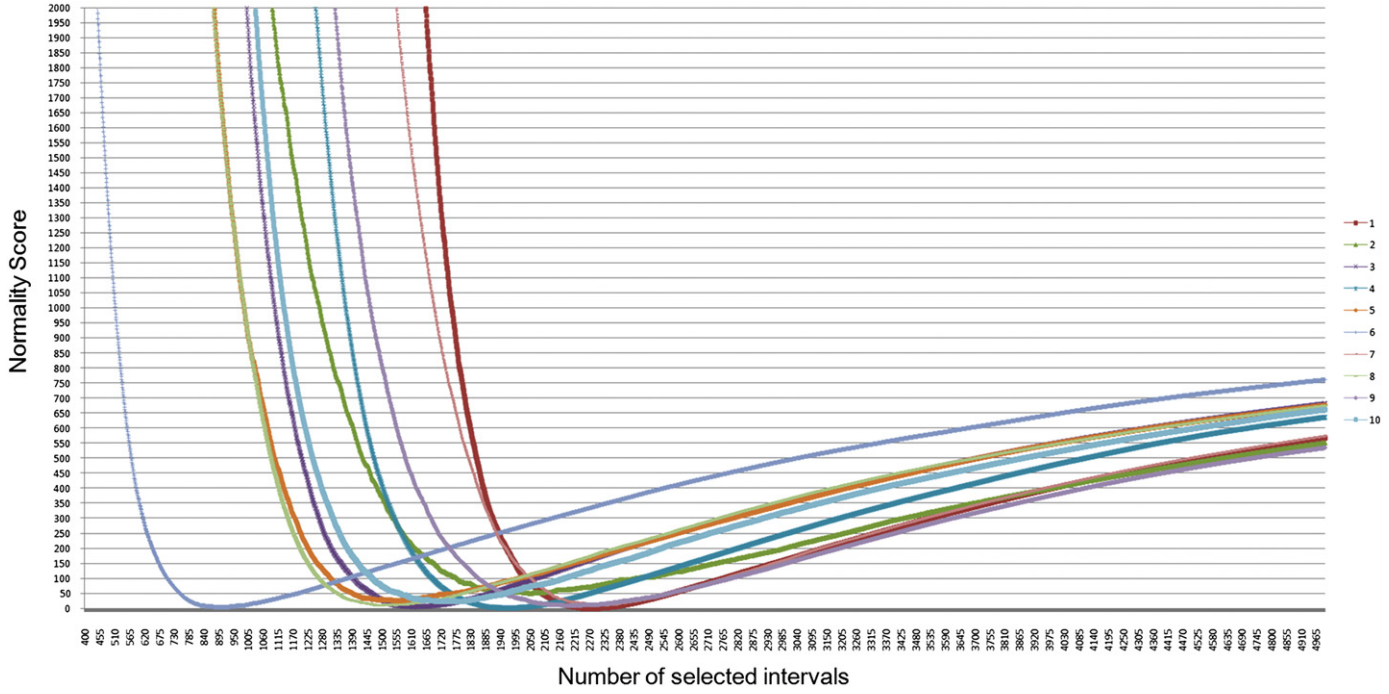

Fig. 2. Changes in the normality score as intervals are removed. Ten randomly chosen samples from chromosome 22 were used. The normality score is zero when the data follow a perfect normal distribution. In all graphs, the normality score rapidly decreases and then slowly increases after a certain point. The intervals selected up until that point are included in CNVs.

genomic variation are removed from the total log ratio values, the normality of the distribution would be reinforced. Specifically, among the various genomic variations, the one that affects the largest number of probes is CNV, because each CNV is generally longer than other structural variations such as SNPs or short repeats. Research on 270 HapMap samples showed that 12% of the whole genome (360 Mb) consists of CNVs [7].

Based on the above discussion, a major component in the creation of a CNV detection algorithm is the selection of probes that are likely to be included in a CNV. More precisely, intervals that are suspected to be a CNV are filtered out, and then a normality test is used to determine whether log ratio values of the remaining intervals follow a normal distribution. This two step process is iterated until the normality score reaches a minimal point (Fig. 2).

To determine the point where the normality score is minimized, intervals with a high average log ratio value are iteratively selected, and the normality score of the remaining intervals is calculated. The Jarque–Bera normality test [10] was used to determine the normality score, and the test statistic *JB* is defined as

$$JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right),$$

where n is the number of observations, and S (*skewness*) and K (*kurtosis*) are defined as

$$S = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2)^{3/2}}, \quad K = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2)^2}$$

The statistic *JB* can be used to test the null hypothesis that the data are normally distributed, which would be indicated by a value of both *skewness* and *kurtosis*. The lower is the *JB* value, the more closely the data fit a normal distribution. Therefore, if *JB* decreases to a certain point and then increases from that point, the selected probes from that point on would not be CNVs.

Table 1
Conditions for the interval *R* being CNV.

If the two intervals of both sides of *R* are not CNVs, *R* is a gain/loss according to the sign of the average log ratio value of *R*.
Else if the two intervals of both sides of *R* are CNVs,
  If the signs of the two intervals' average log ratio values are the same,
    If the sign of the average log ratio value of *R* is the same as the signs of two intervals' average log ratio values, *R* is a gain/loss according to the sign of the average log ratio value.
    Else if the sign of the average log ratio value of *R* is not the same as the signs of two intervals' average log ratio values, *R* cannot be a CNV.
  Else if the signs of the two intervals' average log ratio values are not the same, *R* cannot be a CNV.
Else if the interval of only one side of *R* is CNV,
  If the sign of *R*'s average log ratio value is the same as the sign of the interval's average log ratio value, *R* is a gain/loss according to the sign of the average log ratio value.
  Else if the sign of *R*'s average log ratio value is not the same as the sign of the interval's average log ratio value, *R* cannot be a CNV.

2.4. Postprocessing

The conditions described in Table 1 determine whether or not the selected interval *R* can be a CNV. These conditions are used to form longer CNVs from the selected intervals and also prevent the selection of improper intervals as CNVs.

After all the proper intervals are selected as CNVs, we should find the exact CNVs from them. For all the connected intervals, we find the exact start and end probes that maximize the absolute value of the summed probes in each connected interval.

3. Experimental results

The experiment used 40 samples of high-resolution arrayCGH data published by The Sanger Institute. These data consist of 42 million probes of whole chromosomes, which is approximately

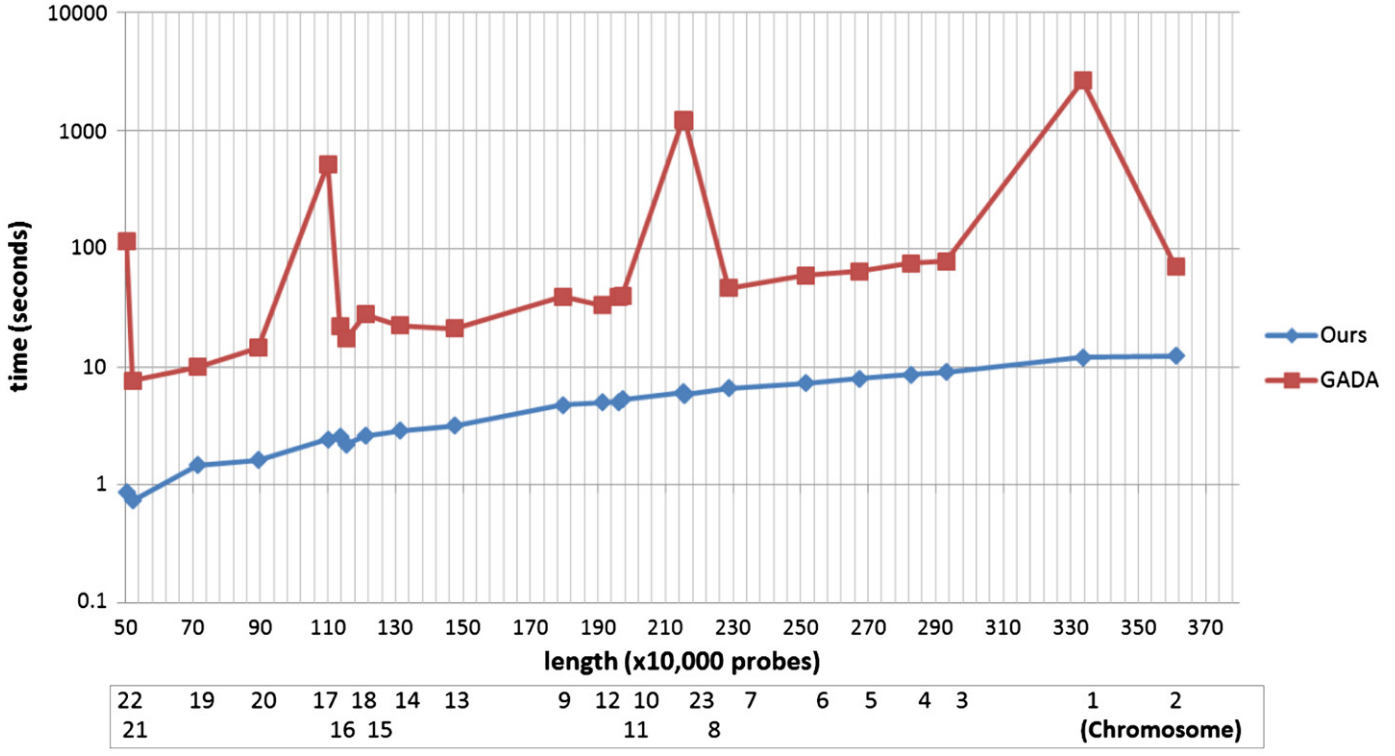

Fig. 3. Average run time of 40 samples for each chromosome.

1600 times as many probes as the 26,574 probes used in a previous study [7].

A total of 763,992,840 base pairs were detected as CNVR, representing 25.28% of the whole genome. The average run time of 40 samples is 115.79 seconds on AMD Athlon X2 Dual, 2.81 GHz, 1.93 GB RAM machine with Windows XP operating system. Fig. 3 shows that ours have near-linear time complexity. The results from this study were compared with the genomic alteration detection algorithm GADA [11] applied by Conrad et al. [12]. A comparison of run-times reveals that GADA runs between 10 and 500 times slower than the algorithm used in this study, and that GADA does not exhibit linear time complexity.

CNVR calls of 40 samples of high-resolution arrayCGH data by Conrad et al. [12] were used to calculate false rates. These calls were based on the definition of a CNVR according to The Sanger Institute, which states that clusters of overlapping CNVs are merged into a CNVR if they have at least 51% of reciprocal overlap. This means that more than two CNVs with similar start and end position can be merged. If two CNVs have similar start points but not similar end points, they cannot be merged into a CNVR. Therefore, CNVs of different sizes can form separate CNVRs, as long as they have at most 50% reciprocal overlap. To construct CNVR, all of the CNVs from 40 samples were organized according to start point. Then, for a given CNV, the remaining CNVs were searched to find those that had more than 51% of overlap with the original CNV, and these overlapping CNVs were merged into one CNVR. The average false positive rate and average false negative rate were 76.52% and 20.02%, respectively. A graphical comparison of the gains, losses, and gain/losses called by Conrad et al. [12] and not by ours, and vice versa, for chromosome 22 is shown in Fig. 4. The figure shows that there are numerous CNVRs not called by Conrad et al., but called by our method, while there are few CNVRs not called by our method, but called by Conrad et al. For other chromosomes, we provide them in tab-delimited text file format as supplementary documents at http://embio.yonsei.ac.kr/~Ahn/cnv.php.

The algorithms from this study were compared with seven algorithms implemented in CGHWeb [13]: CBS [14], FASeg [15], cghFLasso [16], CGHseg [17], Quantreg [18], GLAD [19], and BioHMM [20]. When these seven algorithms were applied to a whole chromosome, they had to be terminated because they were not complete after several days. Therefore, for this study, the algorithms were only run on chromosomes 21 and 22. The results are shown in Table 2.

As indicated in Table 2, the algorithm used in this study had a significant shorter run time than other algorithms. The algorithm from this study processed the 1,024,773 probes up to 1500 times faster than did the standard methods with better accuracy and a relatively low error rate. Although CGHseg and Quantreg have lower false negative rates than the algorithm used in the study,

Fig. 4. Comparison of CNVs called by this study and by Conrad et al. 'Ahn_gain', 'Ahn_loss', and 'Ahn_gain/loss' depict gains, losses, and gain/losses called by our method only, respectively. 'Conrad_gain', 'Conrad_loss', and 'Conrad_gain/loss' depict gains, losses, and gain/losses called by Conrad et al. only, respectively. One small vertical bar in the lower panel indicates one CNV.

Table 2
Benchmark results showing the average run time of 40 samples and the error rates of CNVRs for chromosomes 21 and 22.

Algorithm	Run time (s)	False positive rate (%)	False negative rate (%)
(1) Results for chromosome 21			
This study	0.73	61.83	9.11
Quantreg	28.19	87.33	11.34
cghFLasso	255.16	75.28	30.91
CGHseg	439.86	77.75	1.05
CBS	570.81	13.97	22.72
FASeg	1113.43	47.31	19.02
GLAD	N/A	N/A	N/A
BioHMM	N/A	N/A	N/A
(2) Results for chromosome 22			
This study	0.85	52.59	6.58
Quantreg	25.06	84.92	0.23
cghFLasso	240.26	61.70	9.70
CGHseg	424.72	71.82	4.37
CBS	506.677	23.69	8.68
FASeg	1074.84	47.31	9.73
GLAD	N/A	N/A	N/A
BioHMM	N/A	N/A	N/A

GLAD and BioHMM were terminated after several days wait.

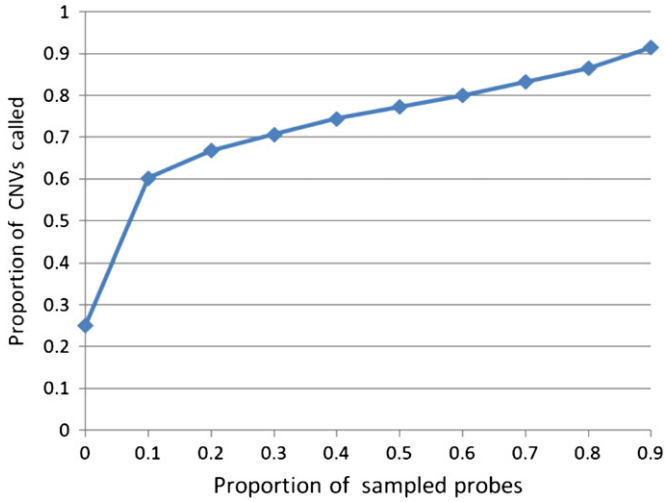

Fig. 5. Average proportion of CNVs called with respect to the proportion of sampled probes.

their false positive rates are much higher. Similarly, although CBS and FASeg have lower false positive rates than the algorithm used in the study, their false negative rates are higher.

While CNVRs are known to cover 12% of the whole genome [7], CNVRs called by The Sanger Institute cover only 4%. This difference guarantees that the CNVRs identified by The Sanger Institute are not comprehensive, and reinforces that 66.67% of false positive rate, at the least, is legitimate. CBS exhibits a notably low false positive rate of 13.97% for chromosome 21 and 23.69% for chromosome 22, indicating that CBS could not detect a comprehensive set of CNVRs in chromosomes 21 and 22.

Finally, an experiment was conducted to evaluate our algorithm's performance with respect to the number of probes covering the CNV. CNVs covered by more than 50 probes for all chromosomes and samples were selected. Then the probes of the selected CNVs were randomly sampled in 10% increments from 90% to 0%, and our algorithm was applied. Fig. 5 shows the average proportion of CNVs that were called. Approximately 60% of the CNVs were detected even when only 10% of the probes were sampled. This indicates that our method is very robust to the number of probes covering the CNV. In addition, approximately 25% of the CNVs were called, even when 0% of the probes were sampled, suggesting that CNVs were removed. Our method can select probes that cover front and rear portions of the removed CNV. Those regions are reported as a CNV, which entirely covers the removed CNV. In this case, the removed CNV is reported to be called by our method.

4. Conclusion

This study proposes a novel CNV detection algorithm using a normality test. Previously established CNV detection methods require parameters that affect the false rates of the results, thus getting optimal parameter values is inevitable. Because there are no precisely validated CNVs, obtaining these optimal parameters is a difficult task. The normality test removes the need to obtain optimal parameters and simply requires that the researcher specify the minimal length of CNV desired, according to the arrayCGH data resolution. Moreover, the algorithm used in this study has a practical runtime even with high-resolution data, as well as low false negative rate. Finally, the algorithm used in this study can be used for arrayCGH data of any resolution.

Conflict of interest statement

None declared.

Acknowledgements

This study makes use of data generated by the Genome Structural Variation Consortium (http://www.sanger.ac.uk/humgen/cnv/42mio/).

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2011-0005154).

References

[1] J.L. Freeman, G.H. Perry, L. Feuk, R. Redon, S.A. McCarroll, D.M. Altshuler, H. Aburatani, K.W. Jones, C. Tyler-Smith, M.E. Hurles, N.P. Carter, S.W. Scherer, C. Lee, Copy number variation: new insights in genome diversity, Genome Res. 16 (8) (2006) 949–961.
[2] L. Feuk, A.R. Carson, S.W. Scherer, Structural variation in the human genome, Nat. Rev. Genet. 7 (2006) 85–97.
[3] S.A. McCarroll, D.M. Altshuler, Copy-number variation and association studies of human disease, Nat. Genet. 39 (2007) S37–S42.
[4] B.E. Stranger, M.S. Forrest, M. Dunning, C.E. Ingle, C. Beazley, N. Thorne, R. Redon, C.P. Bird, A. Grassi, C. Lee, C. Tyler-Smith, N. Carter, S.W. Scherer, S. Tavaré, P. Deloukas, M.E. Hurles, E.T. Dermitzakis, Relative impact of nucleotide and copy number variation on gene expression phenotypes, Science 315 (2007) 848–853.
[5] D.F. Conrad, T.D. Andrews, N.P. Carter, M.E. Hurles, J.K. Pritchard, A high-resolution survey of deletion polymorphism in the human genome, Nat. Genet. 38 (2006) 75–81.
[6] A.J. Iafratem, L. Feuk, M.N. Rivera, M.L. Listewnik, P.K. Donahoe, Y. Qi, S.W. Scherer, C. Lee, Detection of large-scale variation in the human genome, Nat. Genet. 36 (2004) 949–951.
[7] R. Redon, S. Ishikawa, K.R. Fitch, L. Feuk, G.H. Perry, T.D. Andrews, H. Fiegler, M.H. Shapero, A.R. Carson, W. Chen, E.K. Cho, S. Dallaire, J.L. Freeman, J.R. González, M. Gratacòs, J. Huang, D. Kalaitzopoulos, D. Komura, J.R. MacDonald, C.R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M.J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D.F. Conrad, X. Estivill, C. Tyler-Smith, N.P. Carter, H. Aburatani, C. Lee, K.W. Jones, S.W. Scherer, M.E. Hurles, Global variation in copy number in the human genome, Nature 444 (2006) 444–454.
[8] F.N. David, The moments of the z and F distributions, Biometrika 36 (1949) 394–403.
[9] C. Barnes, V. Plagnol, T. Fitzgerald, R. Redon, J. Marchini, D. Clayton, M.E. Hurles, A robust statistical method for case-control association testing with copy number variation, Nat. Genet. 40 (10) (2008) 1245–1252.

[10] C. Jarque, A. Bera, Efficient tests for normality, homoscedasticity and serial independence of regression residuals, Econometric Lett. 6 (1980) 255–259.
[11] R. Pique-Regi, A. Ortega, S. Asgharzadeh, Joint estimation of copy number variation and reference intensities on multiple DNA arrays using GADA, Bioinformatics 25 (2009) 1223–1230.
[12] D.F. Conrad, D. Pinto, R. Redon, L. Feuk, O. Gokcumen, et al., Origins and functional impact of copy number variation in the human genome, Nature 464 (2010) 704–712.
[13] W. Lai, V. Choudhary, P.J. Park, CGHweb: a tool for comparing DNA copy number segmentations from multiple algorithms, Bioinformatics 24 (2008) 1014–1015.
[14] A.B. Olshen, E.S. Venkatraman, R. Lucito, M. Wigler, Circular binary segmentation for the analysis of array-based DNA copy number data, Biostatistics 5 (2004) 557–572.
[15] T. Yu, H. Ye, W. Sun, K. Li, Z. Chen, S. Jacobs, D.K. Bailey, D.T. Wong, X. Zhou, A forward-backward fragment assembling algorithm for the identification of genomic amplification and deletion breakpoints using high-density single nucleotide polymorphism (SNP) array, BMC Bioinf. 8 (2007) 145.
[16] R. Tibshirani, Spatial smoothing and hot spot detection for CGH data using the fused lasso, Biostatistics 9 (2008) 18–29.
[17] F. Picard, S. Robin, M. Lavielle, C. Vaisse, J. Daudin, A statistical approach for array CGH data analysis, BMC Bioinf. 6 (2005) 27.
[18] P.H.C. Eilers, R.X. Menezes, Quantile smoothing of array CGH data, Bioinformatics 21 (2005) 1146–1153.
[19] J.C. Marioni, N.P. Thorne, S. Tavare, BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data, Bioinformatics 22 (2006) 1144–1146.
[20] P. Hupe, N. Stransky, J.P. Thiery, F. Radvanyi, E. Barillot, Analysis of array CGH data: from signal ratio to gain and loss of DNA regions, Bioinformatics 20 (18) (2004) 3413–3422.