HHS Public Access

Author manuscript

J Bioinform Comput Biol. Author manuscript; available in PMC 2015 July 16.

Published in final edited form as:

J Bioinform Comput Biol. 2014 August ; 12(4): 1450021. doi:10.1142/S0219720014500218.

Population clustering based on copy number variations detected from next generation sequencing data

Junbo Duan[*,§], **Ji-Gang Zhang**[†,¶], **Mingxi Wan**[*,‖], **Hong-Wen Deng**[†,‡,**], and **Yu-Ping Wang**[†,‡,††]

[*]Department of Biomedical Engineering, Xi'an Jiaotong University, Xi'an, P. R. China

[†]Department of Biostatistics and Bioinformatics, Tulane University, New Orleans, USA

[‡]Department of Biomedical Engineering, Tulane University, New Orleans, USA

Abstract

Copy number variations (CNVs) can be used as significant bio-markers and next generation sequencing (NGS) provides a high resolution detection of these CNVs. But how to extract features from CNVs and further apply them to genomic studies such as population clustering have become a big challenge. In this paper, we propose a novel method for population clustering based on CNVs from NGS. First, CNVs are extracted from each sample to form a feature matrix. Then, this feature matrix is decomposed into the source matrix and weight matrix with non-negative matrix factorization (NMF). The source matrix consists of common CNVs that are shared by all the samples from the same group, and the weight matrix indicates the corresponding level of CNVs from each sample. Therefore, using NMF of CNVs one can differentiate samples from different ethnic groups, i.e. population clustering. To validate the approach, we applied it to the analysis of both simulation data and two real data set from the 1000 Genomes Project. The results on simulation data demonstrate that the proposed method can recover the true common CNVs with high quality. The results on the first real data analysis show that the proposed method can cluster two family trio with different ancestries into two ethnic groups and the results on the second real data analysis show that the proposed method can be applied to the whole-genome with large sample size consisting of multiple groups. Both results demonstrate the potential of the proposed method for population clustering.

Keywords

Next generation sequencing; copy number variations; non-negative matrix factorization; 1000 Genomes Project

[§]junbo.duan@mail.xjtu.edu.cn
[¶]jzhang9@tulane.edu
[‖]mxwan@mail.xjtu.edu.cn
[**]hdeng2@tulane.edu
[††]wyp@tulane.edu

Competing Interests

The authors declare that they have no competing interests.

1. Introduction

Next generation sequencing (NGS) technology has become the leading platform for genotyping and genomic variation discovery.[1] Unlike traditional technologies such as fluorescence *in situ* hybridization (FISH) and array comparative genomic hybridization (aCGH), NGS provides a direct way to study human genome at the level of base pair, and thus achieved unprecedented resolution. Based on shotgun sequencing, NGS is characterized by its high throughput, enabling output of millions or billions short reads. Recently, various biological and medical studies utilize NGS platforms for de novo assembly,[2] single nucleotide polymorphisms (SNPs) calling,[3] structural variations (SVs) detection,[4] and transcriptome profiling.[5]

Copy number variation (CNV)[6] has been discovered widely in human and other mammal genomes, involving a duplication or deletion of DNA segment of size more than 1 kbp.[7] Similar duplication and deletion events also occur in somatic cells, which are termed copy number alteration (CNA) in oncology. Iafrate *et al*.[8] showed that CNVs are present in human genomes with high frequency (more than 10%). It has been reported that several complex diseases such as autism,[9] schizophrenia,[10] Alzheimer disease,[11] cancer,[12] osteoporosis[13] etc., are associated with CNVs. It is believed that if a CNV region harbors a dosage-sensitive segment, gene expression level varies, and consequently leads to the phenotype abnormality.[14]

Several researches have been carried out for CNV phenotype association[9–12] and CNV detection from NGS data[15–18]; however, the application of CNVs for population study such as the clustering of ethnic groups is still limited. Magi *et al*.[19] showed that samples consisting of two family trios with different ethnicity can be clustered with their CNV profiles. Pearlman *et al*.[20] showed that patients with prostate cancer can be classified into subgroups with different metastatic potential based on their CNA profiles. Their studies suggest that CNV/CNA profile maybe utilized to find bio-markers for group classification or population clustering. Based on their studies, in this paper we show that common CNV, which is a concurrent CNV event occurring at the same genomic location among samples, is a good bio-marker for population clustering.

The proposed population clustering approach based on CNVs consists of five steps: (1) Raw short reads from NGS platform are aligned (or mapped) to the human reference genome (or template, e.g. HG19/NCBI37) with standard alignment tools such as Bowtie[21] or MAQ.[22] (2) Depth of coverage (DOC)[23] or read depth signal is extracted from the alignment data file. (3) The read depth signal is corrected with G-C content.[24] (4) CNVs are detected for each sample with CNV-TV that we recently proposed.[18] (5) Samples are clustered into groups with non-negative matrix factorization (NMF) method[25] based on extracted features. The NMF is a source separation technique,[26] which can cluster common information from multiple data sources.[27] We use the NMF to decompose the matrix consisting of CNVs of all samples into two non-negative matrices, i.e. a source and a weight (or proportion, contribution) matrix. The source matrix contains common CNVs, while the weight matrix shows the contribution or the proportion of common CNVs from each sample, thereby revealing differences between sub-populations.

The paper is organized as follows. First, we outline the models used for CNV detection and population clustering. To evaluate the performance of our method, we apply them to analyze both simulated and two real data set from the 1000 Genome Project. Finally, we discuss the potential application of the method and some open questions.

2. Methods

2.1. CNV detection from single sample

NGS is characterized by shotgun sequencing, which samples reads randomly from the genome. Therefore, the read depth signal obeys the Poisson distribution, whose density parameter (which is equal to the mean and variance) is locally proportional to the copy number. A flat region indicates no CNV event, while a basin or plateau region indicates a copy number deletion or duplication. Mathematically, the detection of CNVs from read depth signal can be formulated as a change-point detection problem.[28] In fact, there are several publicly available CNV detection tools, such as CNV-seq,[17] event-wise testing (EWT),[16] and SegSeq.[15] Since we use CNVs to cluster population, the detection results affect the final output directly. Hence the selection of CNV detection tool should be carefully considered. Based on our comprehensive study of those available tools,[29] we show that total variation (TV) regression based approach, i.e. CNV-TV,[18] achieves more reliable detections with robust performance than several existing methods. As a result, CNV-TV is used as the detection tool. In the following part, we give a brief introduction to CNV-TV.

The CNV-TV model first fits the read depth signal with the TV penalized least squares:

$$\min_{x_i}\left\{\frac{1}{2}\sum_{i=1}^{N}(y_i-x_i)^2+\lambda\sum_{i=1}^{N-1}|x_{i+1}-x_i|\right\}, \quad (1)$$

where N is the length of the read depth signal; y_i,(i = 1,2,…, N) is the read depth signal after G-C content correction; and x_i is the recovered piecewise constant signal. The first term in (1) takes the randomness of read depth into account, and the second term is the TV penalty. Within a region with no copy number changes, $x_i = x_{i+1}$, so no penalty is imposed. At the border between a CNV and non-CNV region there is a change-point, $x_i \neq x_{i+1}$, a penalty $|x_{i+1} - x_i|$ is imposed. λ is the regularization parameter that controls the number of detected CNVs; large λ's yields few detections (e.g. if λ is infinity, $x_1 = x_2 = \cdots = x_N$) and *vice-versa*. The CNV-TV utilizes Schwarz information criterion (SIC)[30] to find optimal parameter.[18]

Once CNVs are detected for a given sample, a feature vector $\boldsymbol{x}$ can be formed as follows: For each CNV region and the rest region (non CNV region), the mean of read depth signal within the region is used. Since different sample may have different coverage, $\boldsymbol{x}$'s are normalized such that non CNV region has same value across samples. In the following part, we consider the population clustering based on the feature vector $\boldsymbol{x}$'s in the context of source separation.[26]

2.2. Population clustering based on CNVs

The source separation techniques first arise in signal processing community. An example is the cocktail party problem, i.e. recovering different speakers' voices from their mixtures

recorded by a set of microphones.[31] The most suitable model for population clustering is the instantaneous mixture,[26] which assumes that a mixture $\boldsymbol{x}_m$ is the weighted-sum of unknown sources:

$$\boldsymbol{x}_m = \sum_{j=1}^{J} w_{jm} \boldsymbol{s}_j, \quad (2)$$

where w_{jm} denotes the weight of the jth source $\boldsymbol{s}_j$ in the mth mixture $\boldsymbol{x}_m$. The corresponding matrix form reads:

$$\boldsymbol{X} = \boldsymbol{S}W, \quad (3)$$

where $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M]$ collects the M mixtures; $\boldsymbol{S} = [\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_J]$ is the J sources; and $\boldsymbol{W} = [w_{jm}]$ is the weight coefficient matrix.

Suppose that a population $\boldsymbol{X}$ contains M samples that derive from J ancestries. By factorizing $\boldsymbol{X}$ into $\boldsymbol{S}$ and $\boldsymbol{W}$, the J ancestries can be recovered as the J columns stored in $\boldsymbol{S}$, and the contribution of each ancestry in the mth mixture sample forms the mth column of weight matrix $\boldsymbol{W}$. Suppose two samples, say the m_1 and m_2th sample, derive from the same ancestry, say the jth source, then the jth entry of the m_1 and m_2th column in $\boldsymbol{W}$ is relatively larger than the rest; therefore, these two columns should be clustered into the same group. In other words, by clustering the columns in $\boldsymbol{W}$, cluster patterns among the samples can be discovered.

It's obvious that the factorizations are not unique and there are infinite number of solutions, some prior knowledge or constraint about matrix $\boldsymbol{S}$ and/or $\boldsymbol{W}$ are required to enforce identifiability, depending on the specific application at hand. In signal processing community, the sources $\boldsymbol{s}_j$'s are usually assumed to be statistically independent, and hence famous algorithms such as independent component analysis (ICA)[32] were proposed. However, ICA may yield negative $\boldsymbol{S}$ and $\boldsymbol{W}$, which make the solution biologically unexplainable with the non-negative NGS read depth data. It makes sense that both $\boldsymbol{S}$ and $\boldsymbol{W}$ should be non-negative matrices, since the former represents the read depth signals of the sources, while the latter represents the weight or contribution of each source. As a results, the NMF approach[25] is a natural choice. Similarly, in image processing and document mining (where the input data matrices are also non-negative as well as the factorization matrices), Lee and Seung demonstrated[27] that NMF can effectively learn common information from the mixtures of patterns. Therefore, by ultilizing NMF, common CNVs are expected to be recovered, which will be used next to discover cluster patterns.

Lee and Seung[27] proposed a multiplicative update algorithm to solve (3):

$$s_{ij} \leftarrow s_{ij} \sum_m \frac{x_{im}}{(\boldsymbol{SW})_{im}} w_{jm}$$
$$w_{jm} \leftarrow w_{jm} \sum_i s_{ij} \frac{x_{im}}{(\boldsymbol{SW})_{im}}.$$

This method is quite easy to implement. However, several works showed that the convergence is not guaranteed.[33] Even though in most cases it converges, the computational speed is low. Therefore, an alternative algorithm based on projected gradient[33] was used for our sequence data.

Finally, we discuss the ambiguity issue in model (3),[25,34] i.e. the factorization cannot be uniquely determined. Even constrains are imposed, there are still several candidate solutions. There are two ambiguities: permutation ambiguity and scale ambiguity. (1) The order of sources is ambiguous up to a permutation. For a given $\boldsymbol{S}$ and $\boldsymbol{W}$, if one permutates the columns of $\boldsymbol{S}$, and permutates the rows of $\boldsymbol{W}$ accordingly, their product does not change at all. (2) The scale (or amplitude) of each source is ambiguous up to a scalar. For a given $\boldsymbol{S}$ and $\boldsymbol{W}$, if one multiplies any column of $\boldsymbol{S}$ with a nonzero value, say α, while divides the corresponding row of $\boldsymbol{W}$ by α, the product keeps the same. These two ambiguities will be demonstrated in the simulation. As a result, further constraints should be imposed to have a unique decomposition; for example, the median of each source is required to be a predefined value. However, since the ambiguity issue does not affect the clustering results and the common CNV discovery, we do not impose any constraint.

3. Results

3.1. Simulation

In the simulation study, two genomes of size 2 Mbp were first simulated, and then three CNVs were artificially introduced into each genome with random size between 3 kbp and 200 kbp. Each CNV has an equal chance to be a homozygous deletion, heterozygous deletion or heterozygous duplication, corresponding to copy number 0, 1, and 3 respectively. We only consider these three cases since these CNVs are both most common and difficult to detect.[35] The read depth signals $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ were simulated (see Fig. 1) such that each has the normal (corresponding to copy number 2) read depth of 200 reads per 1 kbp on average. Then following the mixture model (3), 6 mixed read depth signals $\boldsymbol{y}_i$,(i = 1,2,…,6) were generated. These six samples were divided into two groups (see Fig. 2). For the first three samples, the weight of the first source w_{1m}, (m = 1, 2, 3) obeys the uniform distribution at interval [0.5,1], while for the second group, w_{1m}, (m = 4, 5, 6) obeys the uniform distribution at interval [0, 0.5]. The weight of the second source w_{2m}, (m = 1, 2,…, 6) is $1 - w_{1m}$ such that the sum of weights for a sample is equal to 1. To take the sequencing error into account, a random noise following Poisson distribution with variance 20 (representing 10% unmapped reads) was subtracted from the mixed read depth signals. Figure 2 displays an example of the simulated data set.

After each of the 6 read depth signals was processed by the CNV-TV,[18] CNVs were obtained. Then a data matrix $\boldsymbol{X}$ of size 2000 × 6 was constructed. Afterwards the NMF code written by Lin[33] was used to decompose $\boldsymbol{X}$ into the source matrix $\boldsymbol{S}$ of size 2000 × 2 and weight matrix $\boldsymbol{W}$ of size 2 × 6. The algorithm was initialized with a random positive matrix $\boldsymbol{S}_0$ and $\boldsymbol{W}_0$. Tolerance, limit of time, and maximal iteration number were set to 1e—3, 30, and 1e3, respectively. Figure 3 shows the first and second column of $\boldsymbol{S}$, which are the estimates of $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$. Since the NMF has permutation ambiguity, the first column of $\boldsymbol{S}$ after NMF corresponds to the estimate of the first source in some cases or the second source in

other cases. The same situation occurs for the second column of $\boldsymbol{S}$. Since each mixture includes both sources, there are leaks between the source estimates; e.g. at the location 1500 kbp, the first estimate has a small peak which is from the second estimate at the same location. The clustering result of the columns of weight matrix $\boldsymbol{W}$ is displayed in Fig. 4. It can be seen that the two groups are clearly separated.

Considering the scale ambiguity, to measure how close an estimate is to its real one, the Pearson correlation was calculated. For each source estimate (a column of $\boldsymbol{S}$ after NMF), the Pearson correlations between the estimate with both $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ were calculated, and displayed in Fig. 5. If $\hat{\boldsymbol{s}}$ is a good estimate of $\boldsymbol{s}_1$, the correlation should be high (close to 1), and the correlation with $\boldsymbol{s}_2$ should be low (close to 0). As shown in Fig. 5 with 100 random simulated data set (blue dots and red circles represent the first and second source estimate respectively), there are two clusters around (0,1) and (1,0), indicating that the estimates are highly consistent with real ones. Note that the blue dots and red circles distribute evenly due to the permutation ambiguity. Furthermore, the correlation between the estimated and real weights w always reaches as high as 0.99 and above.

It is reported that NMF is sensitive to the initialization,[36] i.e. the factorization results changes with different initialization of $\boldsymbol{S}_0$ and $\boldsymbol{W}_0$. So we studied whether this affects the clustering performance. We used NMF to decompose a same data set with different initialization strategies reported by Langville *et al*.,[36] and the results show that random initialization is the best for our problem. The correlations of $\boldsymbol{s}$ are shown in Fig. 6, where 100 different random positive matrices were used as initial matrices $\boldsymbol{S}_0$ and $\boldsymbol{W}_0$. The clustering results of $\boldsymbol{W}$ did not change, indicating the little effect of initialization.

3.2. Real data processing

Two real data sets from the 1000 Genomes Project[37] were analyzed. The first data set is from the family trio project, one of the three pilot studies. In this pilot project, the whole-genomes of two family trios were sequenced, including a CEU (Utah residents with northern and western European ancestry) trio: NA12878-daughter, NA12891-father and NA12892-mother, and a YRI (Yoruba in Ibadan, Nigeria) trio: NA19238-mother, NA19239-father and NA19240-daughter. Since the genomes of the six samples are sequenced with high sequencing coverage (42×), we only use the data from chromosome 21 as a demonstration. The preliminary results were presented in our earlier conference paper.[38] To further test the approach on the whole-genome with larger sample size, the second data set is obtained from the low coverage pilot project, with coverage 2–6 ×. We selected 15 subjects including five CEU samples with Coriell ID NA12004, NA12006, NA12044, NA12156, and NA12287, five YRI samples with Coriell ID NA18505, NA18508, NA18511, NA18517 and NA18523, and five JPY (Japanese in Tokyo, Japan) samples with Coriell ID NA18940, NA18942, NA18943, NA18944, and NA18947. These data are from various sequencing platforms; only the data from the Illumina platform SLX were selected. Since the raw short reads were already mapped to NCBI36 with MAQ,[37] the BAM files were downloaded from the 1000 Genomes Project FTP (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/data/), which store the alignment information.

3.2.1. Family trio data set—First, SAMtools[39] was used to generate the DOC profile from the downloaded BAM file. Since the sequencing coverage is high (42×), the window size was set to 1 kbp to achieve high resolution. The read depth bias is corrected with the G-C content profile by the method of Abyzov *et al*.[24] Then CNV-TV that we proposed[18] was used to detect CNVs. Figure 7 shows the detected CNV regions of the six samples within genomic coordinate 40–46 Mbp. We note that each sample of YRI trio has a CNV near genomic coordinate 44.75 Mbp. Afterward $\boldsymbol{X}$ was built and NMF was utilized for decomposition. Figure 8 displays the hierarchical cluster of $\boldsymbol{W}$, and Fig. 9 displays the common CNVs. Interestingly, Fig. 8 shows that the first source estimate has higher weight in the YRI trio compared with the CEU trio (right half of w_1 is "hotter" than the left half). By comparing the two signals in Fig. 9, we found that the first source estimate has a significant CNV that locates near coordinate 44.75 Mbp; this CNV is a common CNV that can significantly differentiate CEU trio from YRI trio. To further verify this result, the read depth signals of the six individual samples are displayed in Fig. 10. It is clear that all the read depth signals of YRI trio have peaks at location 44.75 Mbp, while those of CEU trio do not. This example demonstrates that the proposed method can better identify common CNVs or differentiate samples from different origins.

It is worthwhile to compare a related work published earlier by Magi *et al*.,[19] in which a method called JointSLM was proposed to detect common CNVs from multiple samples. In that work, the same family trio samples were used to test the performance, and the cluster result was shown in Fig. 4 therein. Compared with their results which was obtained from chromosome 1, our cluster result in Fig. 8 is consistent except that the YRI daughter (NA19240) is closer to her mother (NA19238) than her father (NA19239) in genetics. It was shown that the matrix $\boldsymbol{X}$ containing the CNVs can be used directly for clustering. Compared with their methods, our proposed clustering method permits whole-genome analysis based on the weight matrix $\boldsymbol{W}$ in NMF. Recall that the column number of $\boldsymbol{X}$ is the sample size, and the row number of $\boldsymbol{X}$ is the number of windows along the genome. So, if one clusters the columns of $\boldsymbol{X}$ as JointSLM does, there will be difficulty in running whole-genome analysis, since the columns are very long vectors. By factorizing $\boldsymbol{X}$, the resulting $\boldsymbol{W}$ becomes a small matrix whose column number is the sample size, and the row number is the source number J, bounded by the sample size. Therefore, clustering the columns of $\boldsymbol{W}$ is more feasible even for very large sample size.

3.2.2. Low coverage population data set—For the second data set, since the sequencing coverage is low (2–6×), a nonoverlapping window with a width of 10 kbp is used to guarantee that each read depth signal has median value not lower than 100. To normalize the read depth signal due to different sequencing coverage across samples, each signal was scaled such that the median read depth is 100.

After CNVs are detected with CNV-TV,[18] 21 data matrices are formed, corresponding to 21 chromosomes. Each data matrix has 15 columns, corresponding to the 15 samples. The row number of each data matrix is determined by the length of the chromosome, i.e. the length of chromosome divided by the window size. Then NMF is applied to discover common CNVs by decomposing each data matrix $\boldsymbol{X}$ into $\boldsymbol{S}$ and $\boldsymbol{W}$. As is shown in the first real data analysis, $\boldsymbol{W}$ can be used to cluster samples, and $\boldsymbol{S}$ indicates the common CNV regions that have

similar read depths across samples from the same group, but different read depth across groups.

Since the whole genome is too long, the NMF decomposition is carried out on each chromosome separately. To integrate the whole genome information for a better clustering, a filtering was used to keep only the common CNV regions, where the read depth from the three groups are significantly different (at least one pair *t*-test with *p*-value lower than 1e—3). These regions are listed in supplementary material. Figure 11 displays the clustering of the read depth signals within those regions. It is shown that three group patterns is discovered by integrating the information from the whole genome.

To test whether the detected common CNV regions can be used as bio-markers to classify the three groups of samples, we used the leave one out (LOO) cross validation.[40] In each validation, one sample was taken out from the 15 samples as the independent test data, and the remaining 14 samples were used to train a classifier. Here, we used the sparse representation based classifier that we proposed before.[41] The results show that three groups can be classified without errors by using these CNV regions as bio-markers.

Studies of CNVs with diverse populations have found significant differences in the frequencies of CNVs among distinct ethnic groups.[6,42] In our study, unsupervised hierarchical clustering analysis showed that significant differences exist in terms of CNVs among the three ethnic groups. These results suggested the CNVs can be used as bio-markers to classify the three different ethnic populations. We made a comparison of the identified 73 CNV calls (supplementary material) with those in the Database of Genomic Variants (DGV, http://projects.tcag.ca/variation/), a main repository for DNA CNV. It revealed that 69 of the CNV calls overlap more than 50% with the previously reported CNV regions. Among the identified CNVs, some have been indicated as ethnic specific CNVs. For example, we observed the CNV at region 59.92–60.06 Mbp of chromosome 19, including the loci of killer immunoglobulin-like receptor (KIR) gene family. KIR genes are part of the leukocyte receptor complex (LRC), on chromosome 19q13.4. KIR genes modulate the development and activity of natural killer (NK) and some T-cells through interaction with major histocompatibility complex (MHC) class I receptors. These different KIR loci are highly polymorphic and specific to ethnic groups.[43,44] Our findings may provide a better understanding of genomic differences across ethnic groups in terms of CNVs.

4. Conclusion

We have proposed a method that can cluster human samples of different genetic ethnicity based on their high-throughput sequencing data. The method can be summarized in three steps. In the first feature extraction step, CNVs are extracted from the read depth signal from the raw sequencing data. In the second step, the matrix consisting of the feature is factorized into two non-negative matrices, namely a source matrix and a weight matrix. Finally, the weight matrix can be used to cluster the samples into different ethnic groups and the source matrix can be used to discover common CNVs. We have applied the method to both simulated and real data analysis. We note that only data from the Illumina platform was

tested, but the method is applicable to other NGS platforms. This method can also be extended for other purposes such as subtyping.[45]

There are still two open questions. The first lies in the determination of the source number *J*. This parameter needs to be defined before running NMF. If the cluster number is known in advance, there would be no problem. Otherwise, we propose to first use a large value and then gradually decrease it until a good cluster pattern is found. The second is the choice of the window size when counting the read depth signal. Use of a large window size can improve the reliability of CNV detection, but may miss small yet significant CNV due to low resolution. Therefore, further studies are needed to find a good tradeoff.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was partially supported by the Fundamental Research Funds for the Central Universities, Xi'an Jiaotong University Startup Fund, and NIH grants (R01GM109068, R01MH104680).

Biographies

Junbo Duan received B.S. degree in information engineering and M.S. degree in communication and information system from Xian Jiaotong University, China, in 2004 and 2007, respectively, and the Ph.D. in signal processing from Université Henri Poincaré, France, in 2010. After his graduation, he was a postdoctoral fellow in the Department of Biomedical Engineering and Biostatistics and Bioinformatics at Tulane University, USA, until 2013. He is currently an assistant professor in the Department of Biomedical Engineering at Xi'an Jiaotong University. His major research interests are in probabilistic approaches to inverse problems in biomedical engineering and bioinformatics.

Ji-Gang Zhang received Ph.D. in Animal Genetics and Breeding from China Agricultural University in 2005; M.S. and Bachelor's degrees in Shanxi Agricultural University in 2002 and 1999 respectively. He has published over 30 publications on gene selection and disease classification, CNV detection, gene interaction analysis and integration analysis for multi-omics data. He is a research scientist in the Departments of Biostatistics and Bioinformatics at Tulane University. His research interests include data integration and statistical modeling in high-throughput and high-dimensional data space, such as genomics, transcriptomics, epigenomics, and proteomics data.

Mingxi Wan received B.S. degree in geophysical prospecting in 1982 from Jianghan Petroleum Institute, and M.S. and Ph.D. in biomedical engineering from Xi'an Jiaotong University in 1985 and 1989, respectively. Now, he is a professor and chair of the School of Life Science and Technology of Xi'an Jiaotong University. He was a visiting scholar and adjunct professor at Drexel University and the Pennsylvania State University from 1995 to 1996 and a visiting scholar at the University of California, Davis, from 2001 to 2002. He has authored and coauthored more than 100 publications and three books about medical

ultrasound. He is a recipient of several important awards of the Chinese government and universities. His current research interests are medical ultrasound imaging, especially tissue elasticity imaging, contrast and tissue perfusion evaluation, HIFU, and voice science. He is a member of the IEEE.

Hong-Wen Deng received his bachelor's degree in ecology and environmental biology and master's degree in ecology and entomology from Peking University. He received his master's in mathematical statistics and Ph.D. in quantitative genetics from the University of Oregon. Deng was a postdoctoral fellow in the Human Genetics Center at the University of Texas in Houston where he conducted postdoctoral research in molecular and statistical population/quantitative genetics. He also served as a Hughes Fellow in the Institute of Molecular Biology at the University of Oregon. Deng previously served as professor of medicine and biomedical sciences at Creighton University Medical Center, professor of Orthopedic Surgery and basic medical science and the Franklin D. Dickson/Missouri Endowed Chair in Orthopedic Surgery at the School of Medicine of University of Missouri-Kansas City. He is currently the Chair of Tulane Biostatistics and Bioinformatics department and the Director of Center of Bioinformatics and Genomics. Deng is the holder of multiple NIH R01 awards and recipients of multiple honors for his research. He has widely published over 400 peer-reviewed articles, 10 book chapters, and 3 books. His area of interest is in the genetics of osteoporosis and obesity.

Yu-Ping Wang (SM'2006) received B.S. degree in applied mathematics from Tianjin University, China, in 1990, and M.S. in computational mathematics and Ph.D. in communications and electronic systems from Xi'an Jiaotong University, China, in 1993 and 1996, respectively. After his graduation, he had visited positions at National University of Singapore and Washington University Medical School in St. Louis. From 2000 to 2003, he worked as a senior research engineer at Perceptive Scientific Instruments, Inc., and then Advanced Digital Imaging Research, LLC, Houston, Texas. In the fall of 2003, he returned to academia as an assistant professor of computer science and electrical engineering at the University of Missouri-Kansas City. He is currently an associate professor of Biomedical Engineering and Biostatistics and Bioinformatics at Tulane University and a member of Tulane Center of Bioinformatics and Genomics, Tulane Neuroscience program, and Tulane Cancer Center. His research interests lie in the interdisciplinary biomedical imaging and bioinformatics areas, where he has over 130 peer reviewed publications. He has served on numerous program committees and NSF/NIH review panels. He is an associate editor for several journals including *Journal of Neuroscience Methods* and was a member of Machine Learning for Signal Processing technical committee of the IEEE Signal Processing Society.

References

1. Morgan J. Next gen sequencing survey: What laboratory directors are saying about next generation sequencing, GWAS and stimulus. North Amer Equity Res. 2010
2. Lin Y, Li J, Shen H, Zhang L, Papasian CJ, Deng H-W. Comparative studies of de novo assembly tools for next-generation sequencing technologies. Bioinformatics. 2011; 27(15):2031–2037. [PubMed: 21636596]
3. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from nextgeneration sequencing data. Nat Rev Genet. 2011; 12(6):443–451. [PubMed: 21587300]

4. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nat Meth. 2009; 6:S13–S20.
5. Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10:57–63. [PubMed: 19015660]
6. Redon R, et al. Global variation in copy number in the human genome. Nature. 2006; 444(7118): 444–454. [PubMed: 17122850]
7. Freeman J, et al. Copy number variation: New insights in genome diversity. Genome Res. 2006; 16:949–961. [PubMed: 16809666]
8. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. Detection of large-scale variation in the human genome. Nat Genet. 2004; 36(9):949–951. [PubMed: 15286789]
9. Sebat J, et al. Strong association of *de novo* copy number mutations with autism. Science. 2007; 316:445–449. [PubMed: 17363630]
10. Stefansson H, et al. Large recurrent microdeletions associated with schizophrenia. Nature. 2008; 455:232–236. [PubMed: 18668039]
11. Rovelet-Lecrux A, et al. APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. Nat Genet. 2006; 38(1):24–26. [PubMed: 16369530]
12. Campbell P, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet. 2008; 40:722–729. [PubMed: 18438408]
13. Yang T-L, et al. Genome-wide copy-number-variation study identified a susceptibility gene, UGT2B17, for osteoporosis. Am J Hum Genet. 2008; 83(6):663–674. [PubMed: 18992858]
14. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. Annu Rev Med. 2010; 61:437–455. [PubMed: 20059347]
15. Chiang DY, Getz G, Jaffe DB, O'Kelly MJT, Zhao X, Carter SL, Russ C, Nusbaum C, Meyerson M, Lander ES. High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat Meth. 2009; 6:99–103.
16. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 2009; 19:1586–1592. [PubMed: 19657104]
17. Xie C, Tammi MT. CNV-seq, a new method to detect copy number variation using highthroughput sequencing. BMC Bioinform. 2009; 10:80.
18. Duan J, Zhang J-G, Deng H-W, Wang Y-P. CNV-TV: A robust method to discover copy number variation from short sequencing reads. BMC Bioinform. 2013; 14(150):1–12.
19. Magi A, Benelli M, Yoon S, Roviello F, Torricelli F. Detecting common copy number variants in high-throughput sequencing data by using JointSLM algorithm. Nucleic Acids Res. 2011; 39(10):e65. [PubMed: 21321017]
20. Pearlman A, et al. Clustering-based method for developing a genomic copy number alteration signature for predicting the metastatic potential of prostate cancer. J Probab Stat. 2012; 2012:1–19.
21. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10(3):R25. [PubMed: 19261174]
22. Li H, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009; 25(16): 2078–2079. [PubMed: 19505943]
23. Magi A, et al. Bioinformatics for next generation sequencing data. Genes. 2010; 1:294–307. [PubMed: 24710047]
24. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: An approach to discover, genotype, characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011; 21(6):974–984. [PubMed: 21324876]
25. Berry MW, Browne M, Langville AN, Pauca VP, Plemmons RJ. Algorithms and applications for approximate nonnegative matrix factorization. Computational Statistics and Data Analysis. 2006; 52(1):155–173.
26. O'Grady PD, Pearlmutter BA, Rickard ST. Survey of sparse and non-sparse methods in source separation. Int J Imag Syst Tech, Special Issue on Blind Source Separation and De-convolution in Imaging and Image Processing. 2005; 15(1):18–33.

27. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999; 401(6755):788–791. [PubMed: 10548103]
28. Chen J, Wang Y-P. A statistical change point model approach for the detection of DNA copy number variations in array CGH data. IEEE/ACM Trans Comput Biol Bioinform. 2009; 6:529–541. [PubMed: 19875853]
29. Duan J, Zhang J-G, Deng H-W, Wang Y-P. Comparative studies of copy number variation detection methods for next generation sequencing technologies. Plos One. 2013; 8(3):e59128. [PubMed: 23527109]
30. Schwarz G. Estimating the dimension of a model. Ann Stat. 1978; 6:461–464.
31. Cherry CE. Some experiments on the recognition of speech, with one and with two ears. J Acoust Soc Amer. 1953; 25(5):975–979.
32. Hyvärinen A. Survey on independent component analysis. Neural Comput Surveys. 1999; 2:94–128.
33. Lin C-J. Projected gradient methods for nonnegative matrix factorization. Neural Comput. 2007; 19:2756–2779. [PubMed: 17716011]
34. Hyvärinen A, Oja E. Independent component analysis: Algorithms and applications. Neural Networks. 2000; 13:411–430. [PubMed: 10946390]
35. Klambauer G, et al. cn.MOPS: Mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res. 2012
36. Langville, AN.; Meyer, CD.; Albright, R. 12th ACM SIGKDD International Conf Knowledge Discovery and Data Mining. 2006. Initializations for the nonnegative matrix factorization; p. 1-8.
37. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010; 467(7319):1061–1073. [PubMed: 20981092]
38. Duan, J.; Zhang, J-G.; Deng, H-W.; Wang, Y-P. Detection of common copy number variation with application to population clustering from next generation sequencing data. Int Conf IEEE Engineering in Medicine and Biology Society (EMBC); San Diego, CA, USA. 2012. p. 1246-1249.
39. Li H, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009; 25(16): 2078–2079. [PubMed: 19505943]
40. Efron, B.; Tibshirani, RJ. An Introduction to the Bootstrap. Chapman and Hall/CRC; New York: 1993.
41. Cao H, Duan J, Lin D, Wang Y-P. Sparse representation based clustering for integrated analysis of gene copy number variation and gene expression data. Int J Comput Appl. 2012; 19(2):1–14.
42. Pronold M, Vali M, Pique-Regi R, Asgharzadeh S. Copy number variation signature to predict human ancestry. BMC Bioinform. 2012; 13:336.
43. Jiang W, et al. Copy number variation leads to considerable diversity for B but not A haplotypes of the human KIR genes encoding NK cell receptors. Genome Res. 2012; 22:1845–1854. [PubMed: 22948769]
44. Norman P, et al. Meiotic recombination generates rich diversity in NK cell receptor genes, alleles, haplotypes. Genome Res. 2009; 19:757–769. [PubMed: 19411600]
45. Tang W, Duan J, Zhang J-G, Wang Y-P. Subtyping of gliomaby combining gene expression and CNVs data based on a compressive sensing approach. EURASIP J Bioinform Syst Biol. 2013; 1(2) (in press).

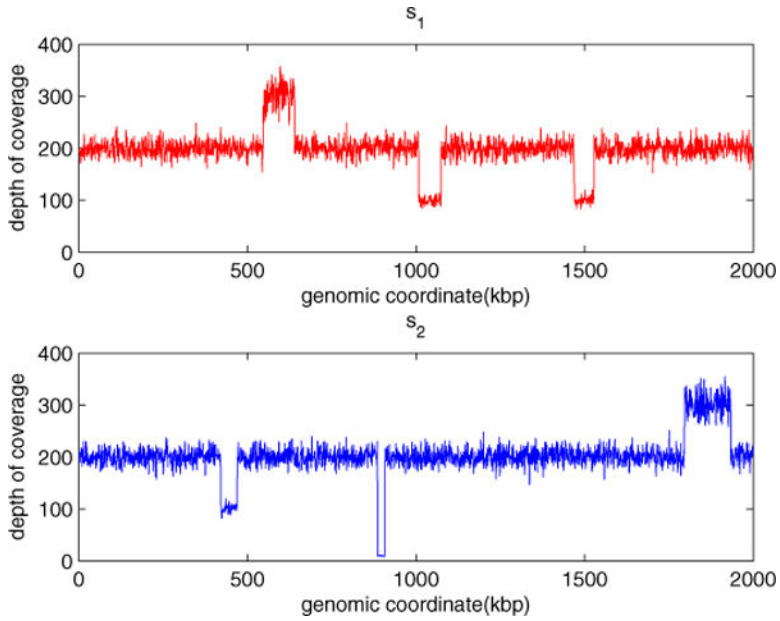

Fig. 1.

An example that demonstrates two source read depth signals $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ in the simulation study.

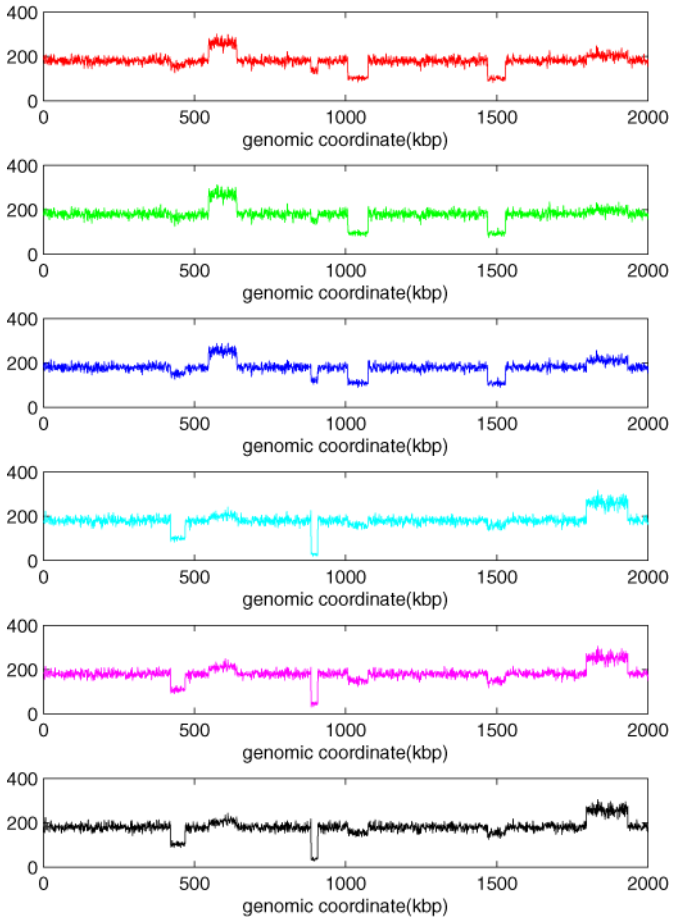

Fig. 2.

The six samples mixed from the two sources displayed in Fig. 1. The first three have larger contribution from the first source than the last three, and therefore they form a group. The last three form another group.

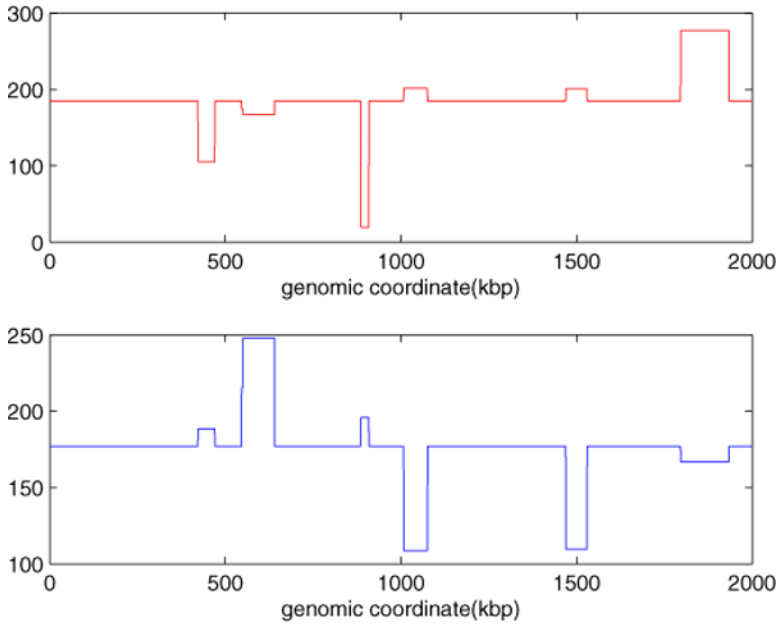

Fig. 3.

The first (upper penal) and second (lower penal) column of source matrix $\boldsymbol{S}$ after the NMF. Note that there is a permutation ambiguity, in fact the first column corresponds to $\boldsymbol{s}_2$, and the second column corresponds to $\boldsymbol{s}_1$. There is also slightly scale ambiguity; note that the base lines are not the same.

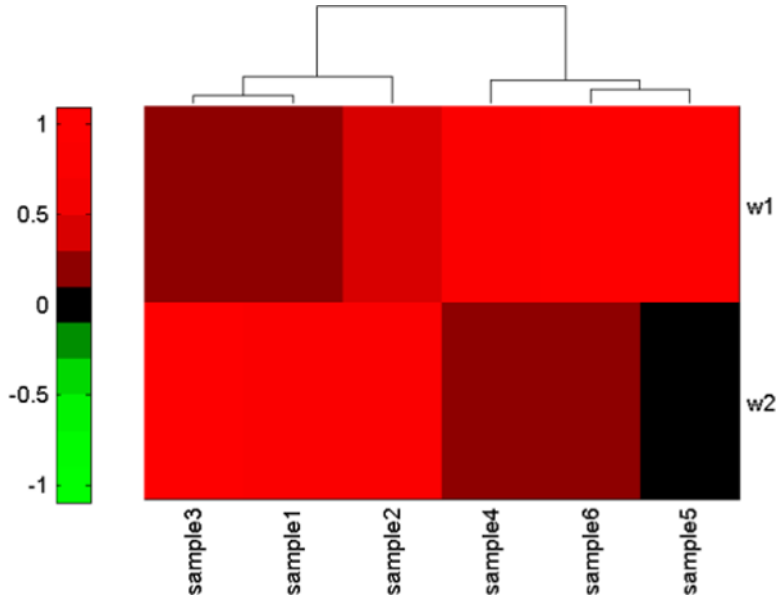

Fig. 4.

Hierarchical cluster of the weight matrix ***W*** of the simulated data set. The two rows labeled *w*1 and *w*2 represent the weights of the two source estimates.

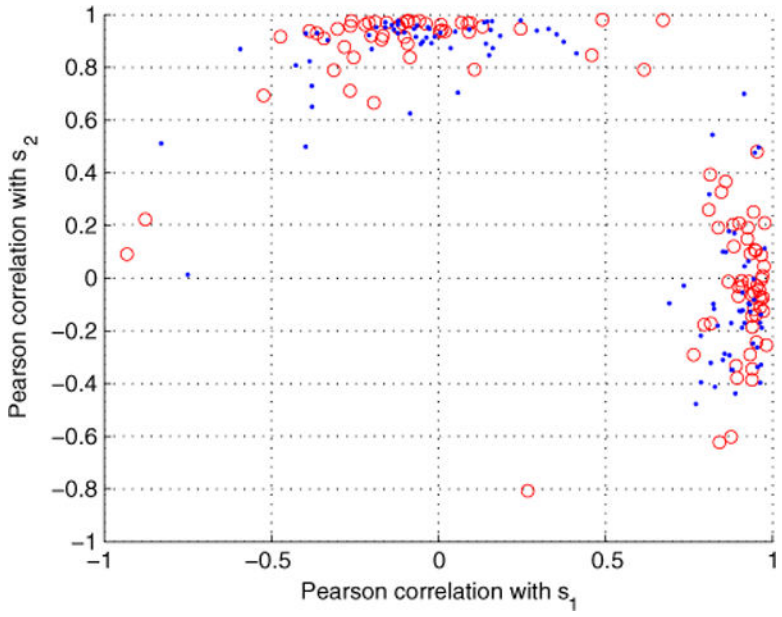

Fig. 5.

The Pearson correlations between the source estimates $\hat{\boldsymbol{s}}_i$,(i = 1,2) and original sources $\boldsymbol{s}_i$,(i = 1,2). Red circles/Blue dots represent the first/second column in source estimate matrix $\hat{\boldsymbol{S}}$, namely, $\hat{\boldsymbol{s}}_1/\hat{\boldsymbol{s}}_2$. Note that since there is a permutation ambiguity, both red circles and blue dots distribute evenly at the two clusters.

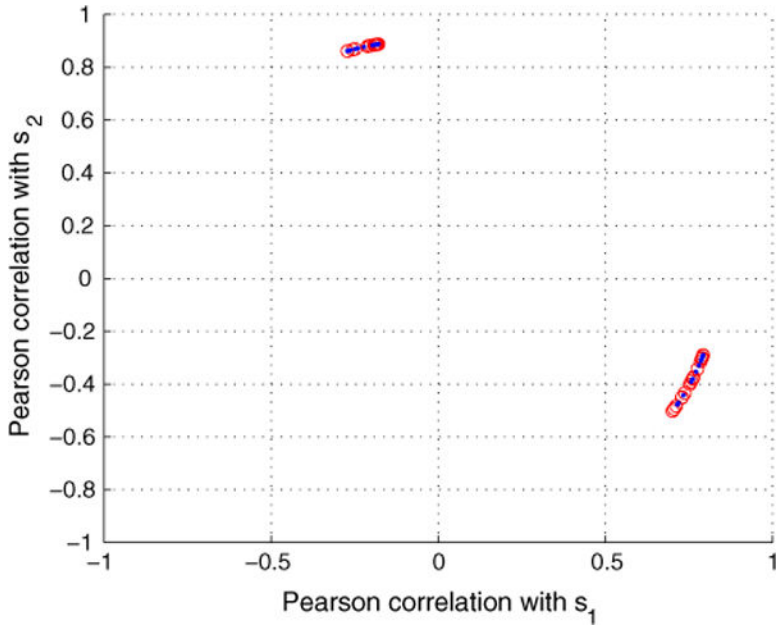

Fig. 6.

The Pearson correlation display as Fig. 5. This figure shows the effect of initialization. The same data set was used but with 100 different random initialization of $\boldsymbol{W}_0$ and $\boldsymbol{S}_0$.

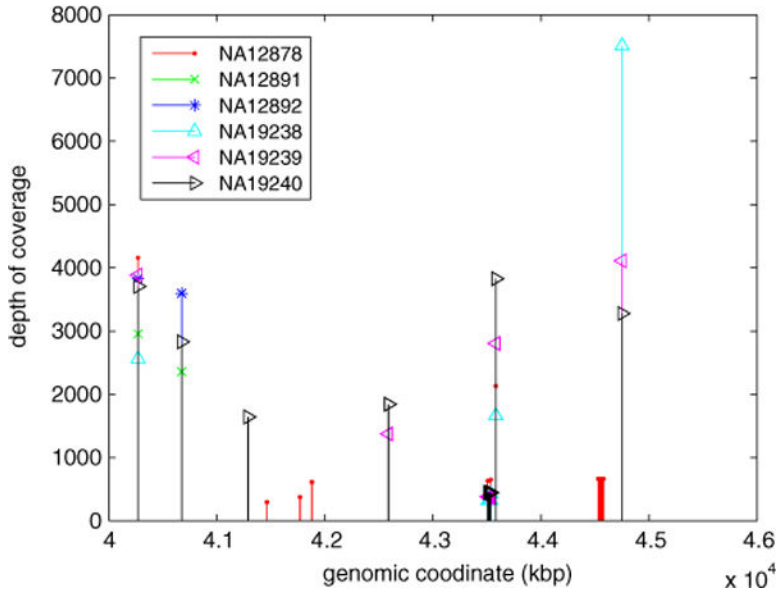

Fig. 7.

Detected CNV regions within 40–46 Mbp. The amplitude of each spike represents the DOC value.

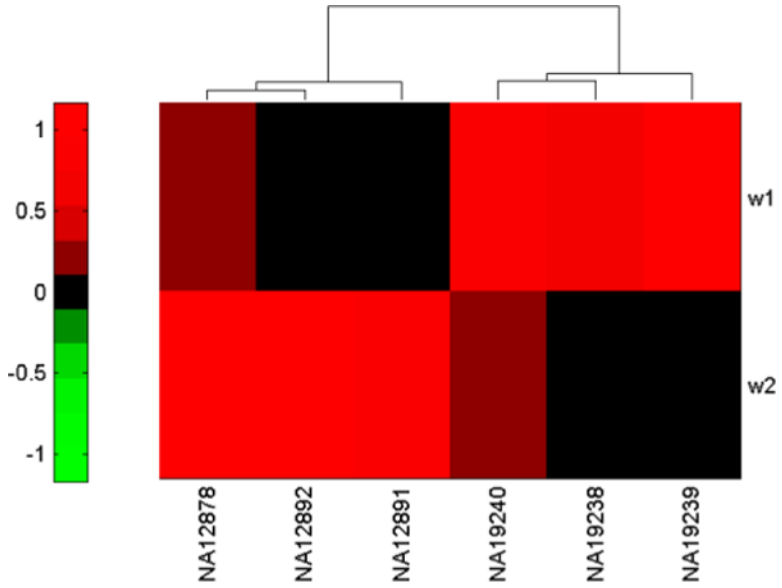

Fig. 8.
Hierarchical cluster of the contribution matrix ***W*** of the first read data set.

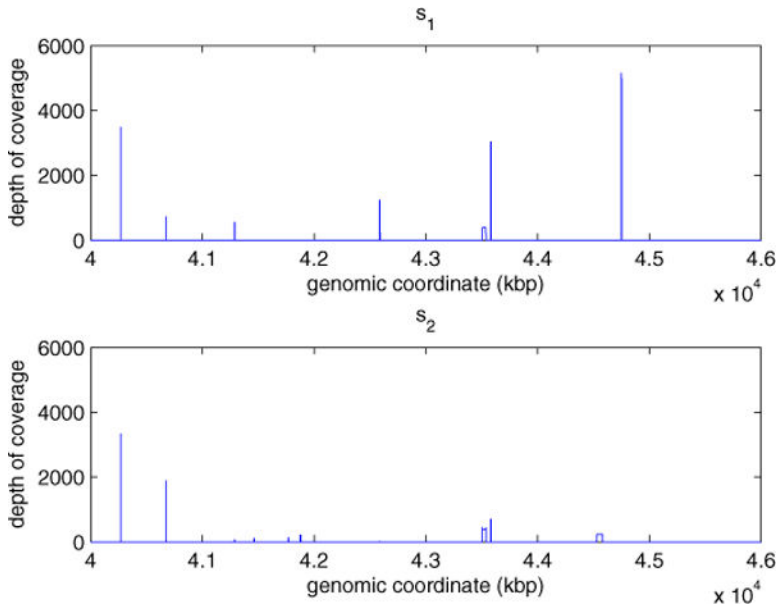

Fig. 9.

The detected common CNV regions of the first/second column (upper/lower penal) of source matrix ***S*** within 40–46 Mbp.

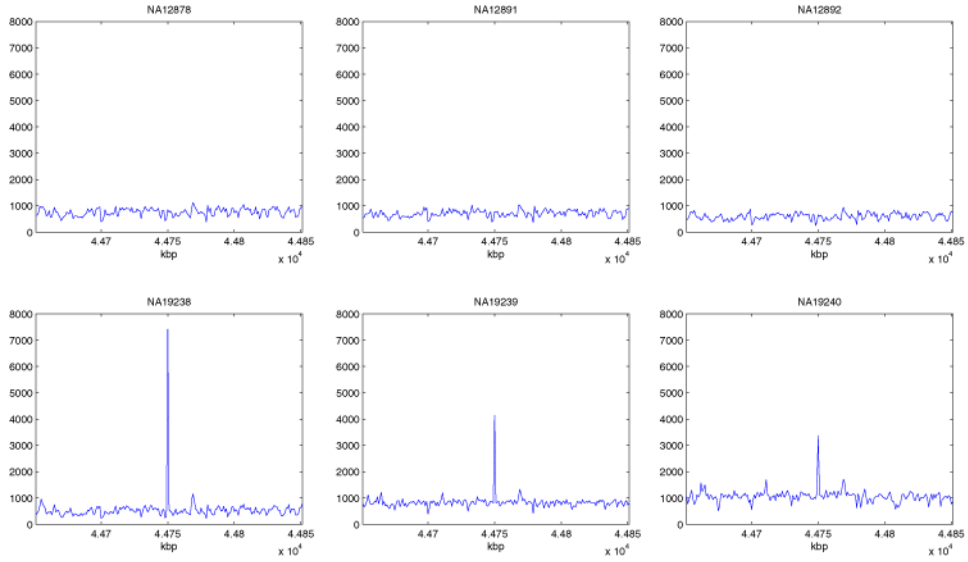

Fig. 10.
The read depth signals of six individual samples within 40–46 Mbp.

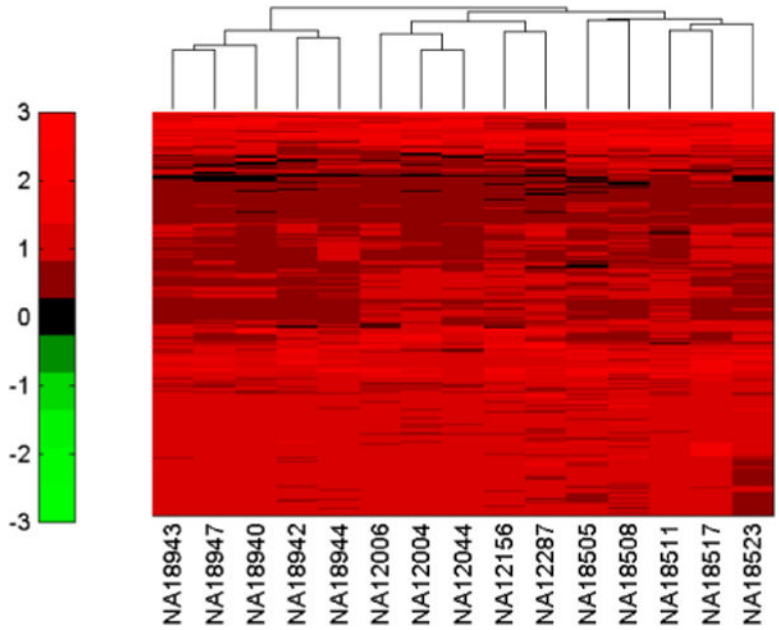

Fig. 11.
Hierarchical cluster of the whole genome of the second real data set.