34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

# Detection of common copy number variation with application to population clustering from next generation sequencing data*

Junbo Duan†, Ji-Gang Zhang‡, Hong-Wen Deng‡ and Yu-Ping Wang†,‡

*Abstract*— **Copy number variation (CNV) is a structural variation in human genome that has been associated with many complex diseases. In this paper we present a method to detect common copy number variation from next generation sequencing data. First, copy number variations are detected from each individual sample, which is formulated as a total variation penalized least square problem. Second, the common copy number discovery from multiple samples is obtained using source separation techniques such as the non-negative matrix factorization (NMF). Finally, the method is applied to population clustering. The results on real data analysis show that two family trio with different ancestries can be clustered into two ethnic groups based on their common CNVs, demonstrating the potential of the proposed method for application to population genetics.**

## I. Introduction

Next generation sequencing (NGS) technology provides a direct way to study human genome in the level of base pair, and thus has received widespread attention in biomedical applications within recent years. Unlike traditional technologies such as fluorescence *in situ* hybridization (FISH) and array comparative genomic hybridization (aCGH), NGS is an high throughput technology that can output million or billion short reads from the shotgun sequencing, and thus provides high resolution mapping of genomic regions. The huge amount of data can be utilized for *de novo* assembly [1], single nucleotide polymorphisms (SNPs) calling [2], structural variations (SVs) detection [3], *etc.*

We focus on the detection of copy number variation (CNV) [4], which covers approximately 10% of human genome. CNV, as a major form of SV, has been associated with complex diseases such as autism [5], schizophrenia [6], Alzheimer disease [7], cancer [8], *etc.* CNVs are the duplication or deletion events of DNA segments with size more than 1 kbp [9]. There have existed several CNV detection methods [10], [11], [12], [13]; however, all of them focus on CNV detection from an individual sample, or two samples including a case and a control sample. In this paper, we consider the detection of common CNVs, which are the recurrent CNVs among a population. These common CNVs can be used for population clustering.

First, the method that was presented in [14] is used to detect CNVs from depth of coverage (DOC) of each sample. Then the non-negative matrix factorization (NMF) method [15] is employed to detect common CNVs. NMF is one of the source separation techniques [16]. Lee and Seung [17] showed that NMF can learn the common information from multiple data sources, which motivated us to apply the proposed method to detect common CNVs. The NMF models the data (detected CNVs in our problem) as the product of a source matrix and a contribution (or weight) matrix. Both of them are non-negative matrices. The source matrix includes the common CNVs, while the contribution matrix includes the weights of common CNVs in each sample. Therefore, using the contribution matrix can cluster population samples into different ethnic groups.

This paper is organized as follows: in Sec. II, the method to detect common CNVs is presented, based on total variation penalized least square optimization and NMF method. In Sec. III, the presented methods are used to population clustering. We processed a data set downloaded from the 1000 Genome Project for catalog of human genetic variations (www.1000genomes.org). The data set includes a CEU trio of European ancestry and a YRI trio of Yoruba Nigerian ancestry, which can be successfully classified based on their CNVs using our proposed approach. The paper is concluded in Sec. IV.

## II. Methods

### A. Copy number variation detection from single sample

The raw NGS data contains a huge amount of short reads. To detect CNVs, firstly these reads need to be mapped to the reference genome, *e.g.* build37 (or hg19) of human. After mapping, we can obtain the depth of coverage (DOC) by counting the number of mapped reads in the fixed-size, non-overlapping and consecutive windows [11]. Because of the correlation between G-C content and DOC [18], G-C content correction on DOC is often needed.

The data we start with are DOC $y_i, (i = 1, 2, \ldots, N)$, where $N$ is the number of windows. Because shotgun sequencing samples reads randomly on the genomic loci, the DOC is locally proportional to the copy number, so flat regions correspond to the same copy number. The detection of CNVs from DOC is modeled as a change-point detection problem, with the basic assumption that $y_i$ is piecewise constant, and the basins/plateaus in $y_i$ correspond to deletions/duplications. Consequently, the CNV detection is formulated as the following total variation penalized least-

*This work was partially supported by NSF and NIH grant.

†J. Duan and Y.-P. Wang are with the Department of Biomedical Engineering, Tulane University, New Orleans, USA. jduan@tulane.edu, wyp@tulane.edu

‡J.-G. Zhang, H.-W. Deng and Y.-P. Wang are with the Department of Biostatistics and Bioinformatics, Tulane University, New Orleans, USA. jzhang9@tulane.edu, hdeng2@tulane.edu

978-1-4577-1787-1/12/$26.00 ©2012 IEEE

square optimization problem:

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

where $x_i$ is the denoised or smoothed version of $y_i$ that can be used to call CNVs. The first term in (1) is the fitting error, and the second term is the total variation penalty. When a change-point presents between $x_i$ and $x_{i+1}$, a penalty $|x_{i+1} - x_i|$, *i.e.* the absolute value of $x_{i+1} - x_i$, is imposed. $\lambda$ is the regularization parameter, which can control the tradeoff between fitting error and penalty caused by change-points. Large $\lambda$ yields low deviation of $x_i$, thus low false positive rate but at the cost of low true positive rate, and *vice-versa*. Because of the page limit, the reader is referred to our earlier work [14] for the detailed explanation of this criterion, the efficient algorithm to solve this problem, and the strategy to select the regularization parameter $\lambda$.

### B. Common copy number variation call

The common CNV detection is considered in the context of source separation. The source separation aims to extract individual sources from their linear or nonlinear mixtures. The most suitable model for our problem is the instantaneous mixture [16], which models the data $\boldsymbol{x}_m \in \boldsymbol{R}^N$ as the weighted-sum of sources:

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

where $w_{jm}$ denotes the contribution of the $j$-th source $\boldsymbol{s}_j$ in the $m$-th mixture data $\boldsymbol{x}_m$. This can be written in matrix form as:

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

where $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M] \in \boldsymbol{R}^{N \times M}$ contains the $M$ mixture data; $\boldsymbol{S} = [\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_J] \in \boldsymbol{R}^{N \times J}$ contains the $J$ sources; and $\boldsymbol{W} = [w_{jm}] \in \boldsymbol{R}^{J \times M}$ is the contribution matrix.

Suppose a population contains $M$ samples $\boldsymbol{X}$ that derive from $J$ ethnic groups $\boldsymbol{S}$, the model (3) characterizes the blood mixing procedure. By factorizing $\boldsymbol{X}$ into $\boldsymbol{S}$ and $\boldsymbol{W}$, the pure blood can be found as the $J$ sources, and the weights of each pure blood in the mix blood form the contribution matrix $\boldsymbol{W}$, which can be further used for population clustering systematically.

When $\boldsymbol{s}_j$'s are assumed statistically independent, famous algorithms like independent component analysis (ICA) [19] can be employed to estimate source matrix $\boldsymbol{S}$ and contribution matrix $\boldsymbol{W}$. However, ICA may yield negative $\boldsymbol{S}$ and $\boldsymbol{W}$, which is mathematically sound but not biologically meaningful. As another approach, given non-negative matrix $\boldsymbol{X}$ and non-negative constraint on both $\boldsymbol{S}$ and $\boldsymbol{W}$, the factorization (3) is known as non-negative matrix factorization (NMF) [15]. From the applications of NMF in the image processing and documents mining, Lee and Seung [17] showed that NMF can learn the common information from the mixture data. Based on this property of NMF, the common CNVs can be found from those detected from each individual sample.

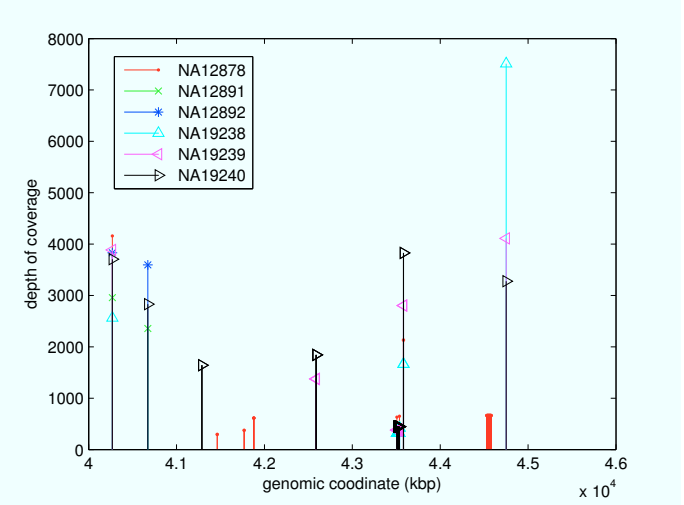


Fig. 1. Detected CNV regions within 40∼46 Mbp. The amplitude of each spike represents the DOC value.

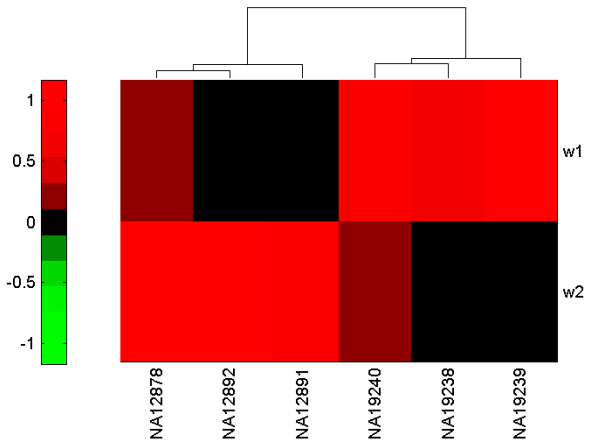


Fig. 2. Cluster of the contribution matrix $\boldsymbol{W}$. The two rows labeled w1 and w2 represent the weights of sources $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$.

Lee and Seung [17] proposed a multiplicative update algorithm to solve (3):

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

which is simple to implement. However, the convergence of this method is not fast enough for our sequence data which has high dimensionality. Therefore, an alternative algorithm based on projected gradient [20] was used in our study.

## III. Results

We downloaded the aligned sequencing data (BAM file) of chromosome 21 of six samples from the 1000 Genomes Project. These six samples include a CEU trio of European ancestry (NA12878-daughter, NA12891-father and NA12892-mother) and a YRI trio of Yoruba Nigerian ethnicity (NA19238-mother, NA19239-father and NA19240-daughter).

For each individual sample, first SAMtools [21] was used to generate the DOC profile from the downloaded BAM file. The window size was set to 1 kbp to reduce the computational burden. Then the method proposed in Sec. II-A was used to detect CNVs. The lower and upper threshold to call a CNV were determined from the histogram of the DOC such that 10% (as CNVs cover approximately 10% of human genome) of short reads falling outside the normal region. The normal region is defined as the interval between the upper and lower threshold with center locations at the peak of the histogram. Fig. 1 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.

Once the CNVs of each individual sample are detected, the DOC of CNV regions were input as the columns of mixture matrix $\boldsymbol{X}$. Each column corresponds to a sample. Regions without CNV are set to 0. Then we used the NMF code written by Lin [20] to factorize $\boldsymbol{X}$ into $\boldsymbol{S}$ and $\boldsymbol{W}$. The algorithm was initialized with random positive matric $\boldsymbol{S}_0$ and $\boldsymbol{W}_0$. Since there are two ethnic groups, the parameter $J$ is set to 2. Fig. 2 displays the hierarchical cluster of $\boldsymbol{W}$, and Fig. 3 displays the two columns of $\boldsymbol{S}$. The cluster result is consistent with that of Magi *et al.* [22], which was obtained from chromosome 1, except that the YRI daughter is genomically closer to her mother than her father. Interestingly, Fig. 2 shows that source $\boldsymbol{s}_1$ (first column in $\boldsymbol{S}$) has higher contribution in the YRI trio compared with the CEU trio (right half of $\boldsymbol{w}_1$ is 'hotter' than the left half). By comparing $\boldsymbol{s}_1$ with $\boldsymbol{s}_2$ in Fig. 3, we found that $\boldsymbol{s}_1$ has a significant CNV located near 44.75 Mbp, indicating that this CNV is a common CNV that can significantly differentiate CEU trio and YRI trio. To verify this result, the DOCs of the six individual samples are shown in Fig. 4. It is clear that all the DOCs of YRI trio have peaks at location 44.75 Mbp, while those of CEU trio do not.

## IV. Conclusion

We have proposed a method that can discover common CNVs based on source separation technique (*i.e.*, NMF). It is shown that using information from common CNVs are significant in the clustering of different ethnic groups. Our analysis on real sequencing data from two family trio supported our method and demonstrate the potential of the method in uncovering the genetic causes of the evolution.

The proposed method is not constrained to classify only two ethnic groups as demonstrated in the Results section. The parameter $J$ controls the number of ethic groups. However, for blind clustering, the choose of $J$ remains an open question.

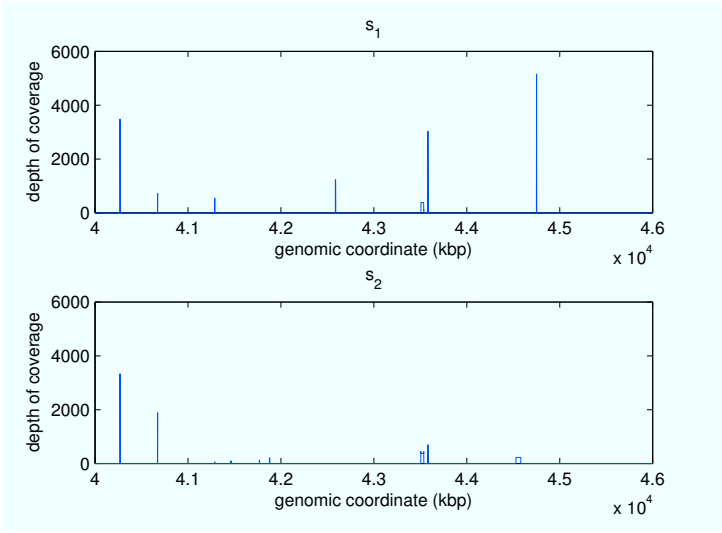


Fig. 3. First/Second column (upper/lower penal) of source matrix $\boldsymbol{S}$ within 40∼46 Mbp.

It's worth noting that two related works were published by Magi *et al* [22] and Klambauer *et al* [23] recently. Similar to our method, both of their proposed methods, namely JointSLM and cn.MOPS, used multiple samples to detect CNVs. But their methods are appropriate under different conditions. The former focuses on the detection of common CNVs that are recurrent at the same location, while the latter intends to significantly reduce the false positives using the information introduced by multiple samples. Magi *et al* also presented the clustering approach based on CNVs. They clustered the columns of mixture matrix $\boldsymbol{X}$ directly. For large sample size, this method is not applicable because of high dimensionality of $\boldsymbol{X}$. Instead, our proposed method cluster the weight matrix $\boldsymbol{W}$, which can not only significantly reduce the data dimensionality but also reduce the variations in the data, resulting in better analysis.

The future studies include the following goals: (1) to employ the proposed method to whole genome analysis. In the current elementary study, to reduce the computation, only the chromosome 21 was processed and displayed, since it is the shortest human chromosome; (2) to compare with other approaches such as JointSLM [22] and cn.MOPS [23]; and (3) to validate the method with more samples in the study of evolutionary genetics.

## References


[1] Y. Lin, J. Li, H. Shen, L. Zhang, C. J. Papasian, and H.-W. Deng, "Comparative studies of *de novo* assembly tools for next-generation sequencing technologies.", *Bioinformatics*, vol. 27, no. 15, pp. 2031–2037, Aug 2011.

[2] R. Nielsen, J. S. Paul, A. Albrechtsen, and Y. S. Song, "Genotype and SNP calling from next-generation sequencing data.", *Nat Rev Genet*, vol. 12, no. 6, pp. 443–451, Jun 2011.

[3] P. Medvedev, M. Stanciu, and M. Brudno, "Computational methods for discovering structural variation with next-generation sequencing", *Nat. Methods*, vol. 6, pp. 13–20, Nov 2009.

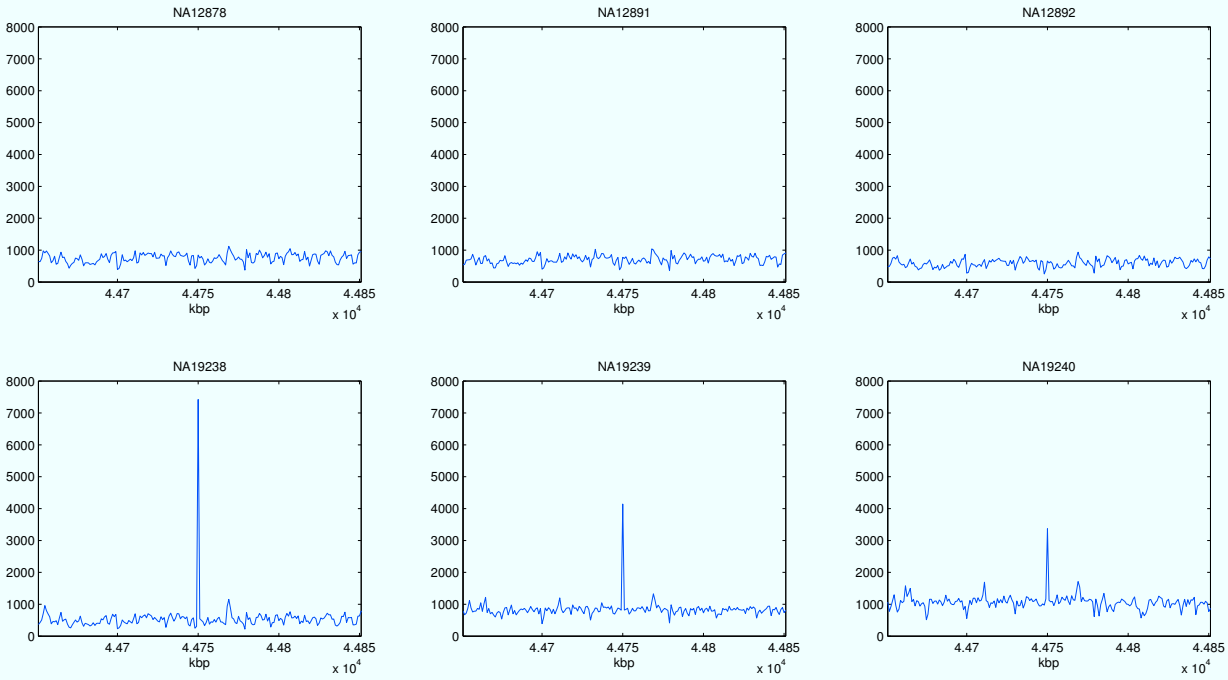


Fig. 4. The DOCs of six individual samples within 40∼46 Mbp.


[4] R. Redon *et al*, "Global variation in copy number in the human genome", *Nature*, vol. 444, no. 7118, pp. 444–454, Nov 2006.

[5] J. Sebat, B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, A. Leotta, D. Pai, R. Zhang, Y.-H. Lee, J. Hicks, S. J. Spence, A. T. Lee, K. Puura, T. Lehtimäki, D. Ledbetter, P. K. Gregersen, J. Bregman, J. S. Sutcliffe, V. Jobanputra, W. Chung, D. Warburton, M.-C. King, D. Skuse, D. H. Geschwind, T. C. Gilliam, K. Ye, and M. Wigler, "Strong association of de novo copy number mutations with autism", *Science*, vol. 316, pp. 445–449, April 2007.

[6] H. Stefansson *et al*, "Large recurrent microdeletions associated with schizophrenia", *Nature*, vol. 455, pp. 232–236, Sep 2008.

[7] A. Rovelet-Lecrux, D. Hannequin, G. Raux, N. L. Meur, A. Laquerrière, A. Vital, C. Dumanchin, S. Feuillette, A. Brice, M. Vercelletto, F. Dubas, T. Frebourg, and D. Campion, "APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy", *Nat Genet*, vol. 38, no. 1, pp. 24–26, Jan 2006.

[8] P. J. Campbell, P. J. Stephens, E. D. Pleasance, S. O'Meara, H. Li, T. Santarius, L. A. Stebbings, C. Leroy, S. Edkins, C. Hardy, J. W. Teague, A. Menzies, I. Goodhead, D. J. Turner, C. M. Clee, M. A. Quail, A. Cox, C. Brown, R. Durbin, M. E. Hurles, P. A. W. Edwards, G. R. Bignell, M. R. Stratton, and P. A. Futreal, "Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing", *Nat. Genet.*, vol. 40, pp. 722–729, Jun 2008.

[9] 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, and C. Lee, "Copy number variation: new insights in genome diversity", *Genome Res.*, vol. 16, pp. 949–961, Aug 2006.

[10] D. Y. Chiang, G. Getz, D. B. Jaffe, M. J. T. O'Kelly, X. Zhao, S. L. Carter, C. Russ, C. Nusbaum, M. Meyerson, and E. S. Lander, "High-resolution mapping of copy-number alterations with massively parallel sequencing", *Nat. Methods*, vol. 6, pp. 99–103, Jan 2009.

[11] S. Yoon, Z. Xuan, V. Makarov, K. Ye, and J. Sebat, "Sensitive and accurate detection of copy number variants using read depth of coverage", *Genome Res.*, vol. 19, pp. 1586–1592, Sep 2009.

[12] C. Xie and M. T. Tammi, "CNV-seq, a new method to detect copy number variation using high-throughput sequencing", *BMC Bioinformatics*, vol. 10, pp. 80, 2009.

[13] J. Duan, J.-G. Zhang, H.-W. Deng, and Y.-P. Wang, "Comparative studies of copy number variation detection methods for next generation sequencing technologies", Tech. Rep., Tulane University, 2012.

[14] J. Duan, J.-G. Zhang, J. Lefante, H.-W. Deng, and Y.-P. Wang, "Detection of copy number variation from next generation sequencing data with total variation penalized least square optimization", in *IEEE international conference on bioinformatics and biomedicine workshops*, Atlanta, GA, USA, Nov. 2011, pp. 3–12.

[15] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization", in *Computational Statistics and Data Analysis*, 2006, pp. 155–173.

[16] P. D. O'Grady, B. A. Pearlmutter, and S. T. Rickard, "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*, vol. 15, no. 1, pp. 18–33, 2005.

[17] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization.", *Nature*, vol. 401, no. 6755, pp. 788–791, October 1999.

[18] D. R. Bentley *et al.*, "Accurate whole human genome sequencing using reversible terminator chemistry", *Nature*, vol. 456, pp. 53–59, Nov 2008.

[19] A. Hyvärinen, "Survey on independent component analysis", *Neural Computing Surveys*, vol. 2, pp. 94–128, 1999.

[20] C.-J. Lin, "Projected gradient methods for nonnegative matrix factorization", *Neural Computation*, vol. 19, pp. 2756–2779, 2007.

[21] H. Li *et al*, "The sequence alignment/map format and SAMtools", *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug 2009.

[22] A. Magi, M. Benelli, S. Yoon, F. Roviello, and F. Torricelli, "Detecting common copy number variants in high-throughput sequencing data by using JointSLM algorithm", *Nucleic Acids Res*, vol. 39, no. 10, pp. e65, May 2011.

[23] G. Klambauer, K. Schwarzbauer, A. Mayr, D.-A. Clevert, A. Mitterecker, U. Bodenhofer, and S. Hochreiter, "cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate", *Nucleic Acids Res*, Feb 2012.