A Sparse Model Based Detection of Copy Number Variations From Exome Sequencing Data

Junbo Duan, *Member, IEEE*, Mingxi Wan, *Member, IEEE*, Hong-Wen Deng, and Yu-Ping Wang*, *Senior Member, IEEE*

Abstract*— *Goal: **Whole-exome sequencing provides a more cost-effective way than whole-genome sequencing for detecting genetic variants, such as copy number variations (CNVs). Although a number of approaches have been proposed to detect CNVs from whole-genome sequencing, a direct adoption of these approaches to whole-exome sequencing will often fail because exons are separately located along a genome. Therefore, an appropriate method is needed to target the specific features of exome sequencing data.** ***Methods:*** **In this paper, a novel sparse model based method is proposed to discover CNVs from multiple exome sequencing data. First, exome sequencing data are represented with a penalized matrix approximation, and technical variability and random sequencing errors are assumed to follow a generalized Gaussian distribution. Second, an iteratively reweighted least squares algorithm is used to estimate the solution.** ***Results:*** **The method is tested and validated on both synthetic and real data, and compared with other approaches including CoNIFER, XHMM, and cn.MOPS. The test demonstrates that the proposed method outperform other approaches.** ***Conclusion:*** **The proposed sparse model can detect CNVs from exome sequencing data with high power and precision.** ***Significance:*** **Sparse model can target the specific features of exome sequencing data. The software codes are freely available at http://www.tulane.edu/ wyp/software/Exon_CNV.m**

***Index Terms*—Copy number variation (CNV), exome sequencing, iteratively reweighted least squares (IRLS), matrix approximation, sparse modeling.**

I. Introduction

THE rapid evolution of next-generation sequencing (NGS) technologies enable us to study genomes with high resolution [1], [2]. As a result, methods aiming to detect genomic variations/mutations based on NGS platforms emerge in the last few years [3], [4]. Among many genomic variations/mutations, copy number variations (CNVs) are extensively studied, which are associated with a variety of complex diseases, *e.g.*, Alzheimer disease [5], autism [6], cancer [7], schizophrenia [8], osteoporosis [9], *etc.* CNV is defined as a type of genomic variation, including duplications/gains or deletions/losses of a DNA segment of size larger than 1 kbp [10]. A widely accepted explanation on the mechanism by which CNVs convey phenotypes is the dosage effect: If a CNV takes place at a genomic region which harbors a dosage-sensitive segment, the corresponding gene expression level increases or decreases depending on the CNV type (duplication or deletion), and consequently leading to the abnormality of phenotype [11].

CNVs present frequently not only in human genome but also in other mammal genomes, so techniques such as multicolor fluorescence *in situ* hybridization (M-FISH) [12], [13] and array comparative genomic hybridization (aCGH) [14] have been applied to their detections. With the emergence of high-resolution NGS, more and more biological investigators migrate to NGS platforms for the study of CNVs, and several detection tools have been developed [15]–[22]. Inspired by the detection methods from aCGH platforms, which use the unbalanced nature of CNVs, the majority of NGS-based methods utilize the read depth signal [4] to detect CNVs. The read depth signal can be viewed as a pileup that bins the aligned short reads into equally sized genomic regions. Therefore, it can reflect the CNV states: The amplified read depths indicate CNV duplications, while the shrunken ones indicate deletions.

To achieve good detection performance in term of sensitivity and specificity, high coverage sequencing is always favorable. However, due to the high cost of sequencing whole genome (about 5000 dollars on July, 2014 [23]), sequencing only the exonic regions was proposed [24]. Exons (or exome) comprise only 1% to 2% of whole human genome [25] and distribute separately, yet they have great importance to genome studies because they are in protein-coding regions that impact gene expression. A direct application of existing CNV tools will often fail to target the specific features of exome sequencing [26], and consequently several methods have been proposed specifically to detect CNVs from exome sequencing [27]–[30]. Earlier studies model read depth signals with Poisson distribution [21], or model ratio of read depth signals with Gaussian distribution [16], [17]. Recent studies show that sophisticated distributions, such as beta binomial [28] or negative binomial distribution [19], [29], are more appropriate. However, for a large pool of data sequenced from several subjects and experiments, data homogeneity could be weak, demanding a more robust model. To deal with this situation, CoNIFER [30] and XHMM [31] utilize principal component analysis (PCA) to reduce the dimensionality of data. From a sparse approximation point of view, the

Manuscript received March 16, 2015; revised June 9, 2015; accepted July 25, 2015. Date of publication August 5, 2015; date of current version February 16, 2016. This work was supported in part by the Fundamental Research Funds for the Central Universities (No. xjj2014060), the China Postdoctoral Science Foundation (No. 2014M560786), the National Natural Science Foundation of China (No. 61401352), the Shaanxi Postdoctoral Science Foundation, and the National Institutes of Health under Grant R01GM109068 and Grant R01MH104680. *Asterisk indicates corresponding author.*

J. Duan and M. Wan are with Xi'an Jiaotong University.

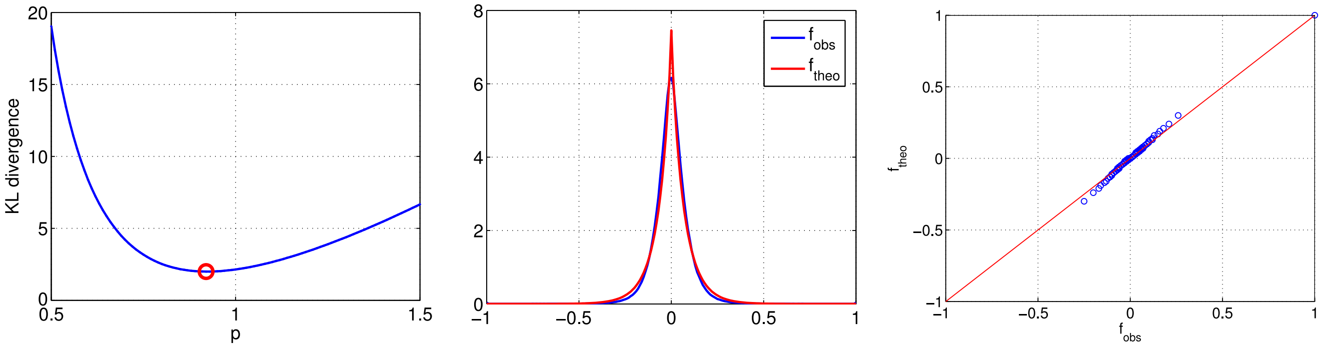
H.-W. Deng is with Tulane University.

*Y.-P. Wang is with the Department of Biomedical Engineering, Department of Biostatistics and Bioinformatics, Tulane University, New Orleans, LA 70118 USA (e-mail: wyp@tulane.edu).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org (File size: 16 KB).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBME.2015.2464674



(a) (b) (c)

Fig. 1. Estimation of parameter p in real data study. (a) KL divergence curve with respect to p ranging from 0.5 to 1.5 with increment of 0.01. The red circle marks the minimum where $p = 0.92$. (b) Observed and theoretical distribution with $p = 0.92$, and the Q–Q plot of these two distributions is displayed in (c).

success of CoNIFER and XHMM is actually because of PCA's ability to capture the sparse structure of sequencing data projected into eigenspace. Motivated by this observation, in this paper, we propose a novel sparse model to detect CNVs from exome sequencing by considering the sparsity of the data.

This paper is organized as follows. In Section II, we introduce a penalized sparse approximation model, and propose an iterative algorithm to optimize the model. Some computational issues will also be discussed in this section. Section III consists of three subsections: In the first subsection, a set of synthetic data is simulated to test the detection performance of the proposed method. In the second subsection, the approach is applied to real data analysis, and its detection performance is compared with three published software, and a known database containing CNVs [32]. In the third subsection, we discuss the performance of the method and some computational issues. This paper is concluded with a summary of the advantage of the proposed method.

II. Methods

A. Notation Conventions

An upper case letter with bold typeface, a lower case letter with bold typeface, and a letter with plain typeface denote a matrix, a column vector, and a scalar, respectively. A letter with a hat denotes the estimate of that variable. For example, $\boldsymbol{A}$ is a matrix, whose jth column is denoted as $\boldsymbol{a}_j$, and the ith element of $\boldsymbol{a}_j$ is denoted as a_{ij}, with estimate $\hat{a}_{ij}$. For matrix $\boldsymbol{A}$, the ℓ_p norm is defined as $\|\boldsymbol{A}\|_p = (\sum_{ij} |a_{ij}|^p)^{1/p}$; and for vector $\boldsymbol{a}_j$, the ℓ_p norm is defined as $\|\boldsymbol{a}_j\|_p = (\sum_i |a_{ij}|^p)^{1/p}$. The ℓ_0 norm is defined as the number of nonzero entries in a matrix or vector. For example, for a scalar a_{ij}, $|a_{ij}|_0 = 0$ if and only if $a_{ij} = 0$; otherwise, $|a_{ij}|_0 = 1$. For vectors and matrices, the entry-wise multiplication (or Hadamard product) and division of two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ with same dimension are denoted as $\boldsymbol{A} \odot \boldsymbol{B} = [a_{ij} b_{ij}]$ and $\boldsymbol{A} \oslash \boldsymbol{B} = [a_{ij}/b_{ij}]$, respectively. For vector $\boldsymbol{a}_j$, $\mathrm{diag}(\boldsymbol{a}_j)$ denotes a diagonal matrix with the main diagonal $\boldsymbol{a}_j$, while for matrix $\boldsymbol{A}$, $\mathrm{diag}(\boldsymbol{A})$ denotes the vector formed by the main diagonal of $\boldsymbol{A}$. $\overline{y_{\cdot n}}$ denotes the mean value of y_{mn} across index m for given n, and the notation of $\overline{y_{m\cdot}}$ is similar.

B. Penalized Sparse Approximation Model

Let y_{mn} denote the read depth at mth exon of nth sample, and let us assume y_{mn} as follows:

$$y_{mn} = d_m v_n c_{mn} + e_{mn} \tag{1}$$

where d_m is the expected normal (without CNV) read depth of the mth exon without coverage difference across samples; v_n is the sequencing coverage of the nth sample; c_{mn} is the copy number state of the nth sample at the mth exon (*e.g.*, 0 for homozygous deletion, 0.5 for heterozygous deletion, 1 for normal, 1.5 for heterozygous duplication, etc.); and e_{mn} is the measurement deviation between the observation and expected value.

Model (1) will be solved with sparse and low-rank approximation [33]. Note that most c_{mn} are equal to 1, so $c_{mn} - 1$ has sparsity. Note also that the normal read depth d_m and sequencing coverage v_n are main factors that contribute to y_{mn}, while c_{mn} and e_{mn} are random variations. The observed data have the degree of freedom of MN since there are MN data points y_{mn}'s. Consider the ideal situation when neither CNV nor measurement deviation presents in the observed data, *i.e.*, $c_{mn} = 1, e_{mn} = 0$, then y_{mn}'s are determined by d_m's and v_n's, whose degree of freedom is $M + N$. So the approximation of y_{mn}'s with d_m's and v_n's has low-rank property. When y_{mn}, d_m, and v_n are expressed in matrix form, this low-rank property is more clear (see Section II-E for more details).

Furthermore, we assume that the normalized measurement deviation e_{mn}/v_n follows identical independent distribution $f(\boldsymbol{\theta})$ with parameter set $\boldsymbol{\theta}$. $f(\boldsymbol{\theta})$ is usually assumed to be Gaussian distribution, but real data analysis shows that Gaussian distribution cannot fit the observation appropriately (see Fig. 1(b), the observed distribution is more spiky compared with a Gaussian distribution). We also simulated dataset with negative binomial distribution, and Kolmogorov–Smirnov test suggested that normalized measurement deviations do not follow the Gaussian distribution (p-value is 5.5e−32). Therefore, we propose to use the generalized Gaussian distribution, which is characterized by one extra parameter [p in (2)] that controls the shape of the distribution.

The zero-center generalized Gaussian distribution reads

$$f(e_{mn}/v_n|p,\sigma) = \frac{p}{2\sigma\Gamma(1/p)} e^{-(|e_{mn}/v_n|/\sigma)^p} \tag{2}$$

where p and σ are shape and scale parameters ($p = 2$ and 1 correspond to the Gaussian and Laplacian distribution, respectively); and $\Gamma(\cdot)$ is the gamma function. The variance of generalized Gaussian distribution is $\delta^2 = \frac{\sigma^2\Gamma(3/p)}{\Gamma(1/p)}$, and therefore, σ can be approximated as

$$\sigma = \delta\sqrt{\frac{\Gamma(1/p)}{\Gamma(3/p)}}. \tag{3}$$

If y_{mn}, p, and σ are known, the maximum likelihood estimator of d_m, v_n, and c_{mn} can be obtained as follows:

$$\max_{d_m, v_n, c_{mn}} \prod_{m,n} f(y_{mn}/v_n - d_m c_{mn}|p,\sigma). \tag{4}$$

Substituting (2) into (4), and taking the negative logarithm, we have

$$\min_{d_m, v_n, c_{mn}} \sigma^{-p} \sum_{m,n} |y_{mn}/v_n - d_m c_{mn}|^p + c(p,\sigma) \tag{5}$$

where $c(p,\sigma)$ is a constant with respect to d_m, v_n, c_{mn} and, therefore, can be omitted in the optimization criterion. Similarly, the coefficient σ^{-p} can be omitted as well.

The estimation of d_m, v_n, and c_{mn} from y_{mn} in (5) is ill-posed since there are more unknown variables than known ones. Based on the Tikhonov regularization framework [34], to address this issue, a penalty term is needed. Note that CNVs cover only a small portion of whole genome (10% [35]); therefore, most loci are at normal status, *i.e.*, most c_{mn}'s equal to 1. We incorporate this prior knowledge into the model by introducing a penalty term $\lambda|c_{mn} - 1|_0$. Here, λ is the penalty level caused by a CNV event, and $|\cdot|_0$ is the sparsity inducing norm that constrains most of c_{mn} to 1 when λ is sufficiently large [36]. When there is a CNV event, *i.e.*, $c_{mn} \neq 1$, a penalty is added to the criterion; otherwise, no penalty is imposed.

Sparse models are originally formulated with ℓ_0 norm, which, however, often lead to a NP-hard problem [37] and become computationally prohibited. Therefore, ℓ_1 norm is employed to relax the problem, which can also enforce sparsity [38] and the corresponding model can be solved more efficiently [39]–[41]. However, the ℓ_1 solution is biased [42]. Therefore, we use ℓ_0 norm in our model, and we show later that we can develop an efficient solver for ℓ_0 norm-based optimization problem.

Based on above analysis, the estimate of d_m, v_n, c_{mn} can be obtained from the following optimization:

$$\min_{d_m, v_n, c_{mn}} \sum_{m,n} \{|y_{mn}/v_n - d_m c_{mn}|^p + \lambda|c_{mn} - 1|_0\}. \tag{6}$$

C. Optimization Algorithm

When $p < 1$ (this is the case in our study, as can be seen in the subsection of real data analysis), the solution to the problem (6) becomes more complicated because of the nonconvexity of both ℓ_p and ℓ_0 norm. To address the optimization problem involving ℓ_p norm with $p < 1$, Daubechies *et al.* proposed an iteratively reweighted least squares (IRLS) algorithm [43]. In this subsection, we propose a two-step optimization procedure based on IRLS.

Because of the binary characteristic of ℓ_0 norm, the penalty term $\lambda|c_{mn} - 1|_0$ equals to either 0 or λ. As a result, if d_m's and v_n's are known, the minimum of (6) occurs at

$$c_{mn} = \tau_\lambda(y_{mn}, d_m, v_n) = \begin{cases} 1, & |y_{mn}/v_n - d_m|^p \leqslant \lambda \\ \dfrac{y_{mn}}{v_n d_m}, & \text{otherwise.} \end{cases} \tag{7}$$

Based on this observation, the first step of the proposed two-step optimization procedure aims to solve (6) with respect to d_m and v_n when all c_{mn}'s equal to 1, or

$$\min_{d_m, v_n} \sum_{m,n} |y_{mn}/v_n - d_m|^p. \tag{8}$$

In the second step, c_{mn}'s are estimated according to (7) given d_m's and v_n's.

To address problem (8), d_m's and v_n's are estimated alternatively: estimate d_m's for fixed v_n's, and then estimate v_n's for fixed d_m's.

1) When fixing v_n's, problem (8) could be decomposed into M subproblems with dimension 1, which reads (for fixed m)

$$\min_{d_m} \sum_n |y_{mn}/v_n - d_m|^p. \tag{9}$$

Since ℓ_p is non-Lipschitz continuous, the minimum of (9) should occur at one of those non-Lipschitz points [44], *i.e.*, y_{mn}/v_n. To find the optimal minimizer, an exhaustive search of all possibilities is needed.

2) When fixing d_m's, problem (8) could be decoupled into N subproblems

$$\min_{v_n} \sum_m |y_{mn}/v_n - d_m|^p. \tag{10}$$

This is a nonconvex criterion since ℓ_p norm is involved. So we adopt the IRLS algorithm [43]. The IRLS iteratively constructs a weight vector $\boldsymbol{w}$ of dimension M with

$$w_m = |d_m - y_{mn}/v_n|^{p-2}. \tag{11}$$

Then, the regression coefficient v_n is updated as

$$v_n = \left(\sum_m (w_m y_{mn}^2)\right)\left(\sum_m (w_m y_{mn} d_m)\right)^{-1}. \tag{12}$$

D. Determination of Tradeoff Parameter λ

The determination of λ is crucial for the final detection performance since λ controls the tradeoff between the data-fitting term and the penalty term [see (6)]. It is obvious that small λ tends to encourage the detection of CNV events among probes, and consequently, both power and false detection rate will increase, and *vice-versa*.

Theoretically, under the Bayesian framework, λ can be calculated if the prior knowledge about σ and c_{mn}'s are given *a priori* [45], but in real applications, these information are

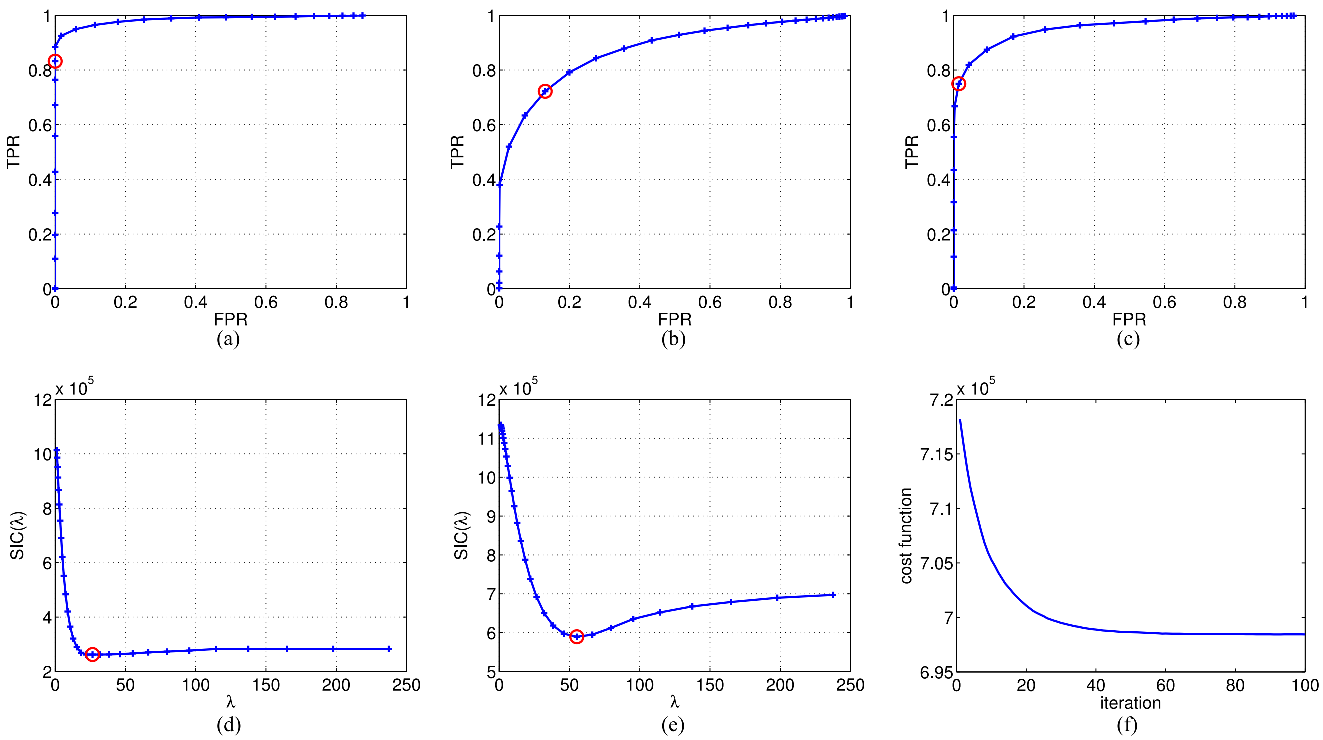


Fig. 2. Results of the proposed algorithm in Simulation I. Blue plus markers represent different level of λ and red circle corresponds to the solution that yields lowest SIC. (a) and (b) ROC of the algorithm for data with generalized Gaussian noise added with variance 25 and 225, respectively. (c) Read depth is sampled with the Poisson distribution. (d) and (e) SIC curves correspond to (a) and (b). The AUC of (a), (b), and (c) are 0.99, 0.90, and 0.95, and the accuracies at the red circles are 0.99, 0.86, and 0.98, respectively. (f) Typical cost function with respect to the iteration index.

not readily available. Empirically, researchers can tune a wide range of λ values manually and make the decision based on the problem at hand. However, an automatic tuning procedure to choose this parameter is desirable. To this end, a model selection technique based on a certain criterion is employed, *e.g.*, the L-curve [46], Akaike information criterion [47], Schwarz information criterion (SIC) [48]. We propose to use the SIC for its robust performance [49], which has been adopted in our earlier CNV detection studies [14], [22] and also in other work [50].

After the iterative optimization procedure, d_m, v_n, c_{mn} are determined for a given λ with (7). The corresponding SIC at this λ is calculated as

$$\mathrm{SIC}(\lambda) = \sum_{m,n} \left\{ \frac{|y_{mn}/v_n - d_m c_{mn}|^p}{\left(\sqrt{\frac{\Gamma(1/p)}{\Gamma(3/p)}}\delta\right)^p} + \ln(MN)\kappa \right\} \tag{13}$$

where the first term is the log-likelihood (δ is the standard deviation defined in Section II-B), and the second term takes into account both the number of observations MN, and model complexity κ, which is the number of free c_{mn}'s (those c_{mn}'s not equal to 1).

One can tune λ on a grid, and the corresponding $\mathrm{SIC}(\lambda)$'s are evaluated according to (13). The best λ is achieved at the point with the lowest SIC value (red circles in Fig. 2(d) and (e)):

$$\hat{\lambda} = \arg\min_{\lambda} \mathrm{SIC}(\lambda). \tag{14}$$

E. Computational Issues

1) Matrix Acceleration: The proposed algorithm was presented for scalar variables. If one implements with scalar operations, the computation burden should be very heavy since there are several nested loops. For efficient implementation, in the following, we reformulate the algorithm for matrix operations.

The matrix alternative of model (1) reads

$$\boldsymbol{Y} = (\boldsymbol{d}\boldsymbol{v}^T) \odot \boldsymbol{C} + \boldsymbol{E} \tag{15}$$

where $\boldsymbol{d}$ collects the expected read depth of M exons; $\boldsymbol{v}$ collects the coverage of N samples; $\boldsymbol{Y} = [y_{mn}]$, $\boldsymbol{C} = [c_{mn}]$, and $\boldsymbol{E} = [e_{mn}]$ denotes read depth data, copy number states, and measurement deviations, respectively. Based on these notations, the matrix form of optimization problem (6) reads

$$\min_{d,v,C} \left\{ \|\boldsymbol{Y}(\mathrm{diag}(\boldsymbol{v}))^{-1} - (\mathrm{diag}(\boldsymbol{d}))\boldsymbol{C}\|_p^p + \lambda\|\boldsymbol{C} - 1\|_0 \right\}$$
$$\text{s.t. } \|\boldsymbol{d}\|_2 = \delta M. \tag{16}$$

Note that in model (15), the outer product between vector $\boldsymbol{d}$ and $\boldsymbol{v}$ has rank 1, and $\boldsymbol{C}$ -1 is a sparse matrix. So this model approximates raw data matrix $\boldsymbol{Y}$ with a low-rank sparse model [33], and considers other factors as variations or noise.

Note also that when no constraint is imposed, the model yields $v_n \to +\infty, d_m \to 0, c_{mn} \to 1$. To avoid this, we impose a constraint on $\boldsymbol{d}$ such that it has a fixed ℓ_2 length δM, where δ is the row standard deviation of $\boldsymbol{Y}$, and is calculated as the mean of the standard deviations of the row vectors of $\boldsymbol{Y}$.

TABLE I
SUMMARY OF ALGORITHM

```
Input: $\boldsymbol{Y}$, p, $\lambda$.

Step0: Initialize $\boldsymbol{v}$.
Step1: Estimate $\boldsymbol{d}$ according to (18),
       normalize $\boldsymbol{d}$: $\boldsymbol{d} = \boldsymbol{d}/\|\boldsymbol{d}\|_2 \delta M$,
       update $\boldsymbol{W}$ according to (19),
       estimate $\boldsymbol{v}$ according to (20),
       iterate Step1 until reaches stopping condition.
Step2: Estimate $\boldsymbol{C}$ according to (17).
Output: $\boldsymbol{v}$, $\boldsymbol{d}$, and $\boldsymbol{C}$.
```

According to (7), if knowing $\boldsymbol{Y}$, $\boldsymbol{d}$, and $\boldsymbol{v}$, the estimate of $\boldsymbol{C}$ is given by

$$\boldsymbol{C} = \tau_\lambda(\boldsymbol{Y}, \boldsymbol{d}, \boldsymbol{v}). \tag{17}$$

If knowing $\boldsymbol{Y}$ and $\boldsymbol{v}$, the estimate of $\boldsymbol{d}$ is

$$\boldsymbol{d} = ES(\boldsymbol{Y}\mathrm{diag}(\boldsymbol{v})^{-1}) \tag{18}$$

where ES performs an exhaustive search of (9) over each row.

Finally, according to (11) and (12), if knowing $\boldsymbol{Y}$ and $\boldsymbol{d}$, the estimate of $\boldsymbol{v}$ reads

$$\boldsymbol{W} = |\boldsymbol{d}\boldsymbol{i}_N^T - \boldsymbol{Y}\mathrm{diag}(\boldsymbol{v})^{-1}|^{p-2} \tag{19}$$

$$\boldsymbol{v} = \mathrm{diag}(\boldsymbol{W}^T(\boldsymbol{Y} \odot \boldsymbol{Y})) \oslash \mathrm{diag}(\boldsymbol{W}^T(\boldsymbol{Y} \odot (\boldsymbol{d}\boldsymbol{i}_N^T))) \tag{20}$$

where $\boldsymbol{i}_N$ is an all-one vector of length N.

2) Initialization and Stopping Condition: To start the iteration, an initial guess of $\boldsymbol{v}$ is needed. Since v_n represents the coverage of the nth sample $\boldsymbol{y}_n$, we use the ℓ_2 norm $\|\boldsymbol{y}_n\|_2$ to initialize v_n. The algorithm converges fast [see Fig. 2(f)], and we terminate the program when the iteration index exceeds a predefined number (*e.g.*, 100).

The quasi-code of the proposed algorithm to optimize problem (16) is shown in Table I.

III. RESULTS AND DISCUSSION

A. Test on Synthetic Data

The test of synthetic data includes two simulations. In the first simulation, the read depth data y_{mn} were generated directly from a statistical model, while in the second simulation, the synthetic data were sampled by following the NGS protocol, *i.e.*, a hybrid model was used.

Since in the simulation studies the ground truth are available, we can evaluate the detection performance by utilizing the receiver operating characteristic (ROC) curve, which displays the true positive rate (TPR or sensitivity/statistical power/recall) versus false positive rate (FPT or 1-specificity). Other statistics can also be deduced from these two quantities. In particular, we will use the area under the curve (AUC) to quantify the detection performance; this quantity takes value between 0 and 1, and for a perfect detector, this quantity should reach to 1.

1) Simulation I: First, the ideal coverage vector $\boldsymbol{v}_0$ and read depth vector $\boldsymbol{d}_0$ were sampled. The column vector $\boldsymbol{v}_0$ has $N = 60$ entries representing 60 samples, and each entry follows a uniform distribution within the interval between 0.9 and 1.0, so the sequencing coverage of 60 samples is appropriate. The column vector $\boldsymbol{d}_0$ consists of $M = 1000$ target probes, and each follows a uniform distribution between 10 and 100. We choose 100 as the upper bound to accord with the order of magnitudes of the real data in the next section, in which 0.67% read depth exceed this bound. We choose uniform distribution to simulate data because this distribution has maximal entropy and does not favor any specific value or distribution. Second, the ideal copy number state matrix $\boldsymbol{C}_0$ with 1000 rows and 60 columns was sampled. Each entry of the matrix has 1% probability to be 0, 4% chance to be 0.5, 90% to be 1, 4% to be 1.5, and 1% to be 2, representing 2-copy (homozygous) deletion, 1-copy (heterozygous) deletion, normal state, 1-copy duplication, and 2-copy duplication, respectively. So totally 10% loci cover CNVs, being consistent with previous study [35]. The noise-free exonic read depth matrix is calculated as $\boldsymbol{Y}_0 = \boldsymbol{d}_0\boldsymbol{v}_0^T \odot \boldsymbol{C}_0$. Two methods were employed to simulate the sequencing uncertainty: in the first method, the independent identically distributed noise was assumed to follow a generalized Gaussian distribution with $p = 0.8, \sigma = 4$ (close to real data) and was added to ideal $\boldsymbol{Y}_0$, yielding the synthetic dataset $\boldsymbol{Y}$; in the second method, each entry of $\boldsymbol{Y}$ follows the Poisson distribution, with parameter being the corresponding entry in $\boldsymbol{Y}_0$.

After the synthetic data were simulated, the proposed detection algorithm was called to detect CNVs. The maximal iteration number was set to 100 for fixed λ, and λ took totally 31 values by following a geometric sequence from $1.2^0 = 1$ to $1.2^{30} = 237.4$ with a common ratio 1.2. Fig. 2 shows the performance of the algorithm in Simulation I, where panels (a) and (b) show the ROC when generalized Gaussian noise with the variance of 25 and 225 was added, respectively, and (c) is the ROC with the Poisson distribution. Each blue dot corresponds to a TPR versus FPR scenario for a given λ. When the tradeoff parameter λ increases, the ROC trajectory evolves from northeast (high TPR and FPR) to southwest region (low TPR and FPR) indicating that a small λ tends to deliver high TPR and *vice versa*. It is also shown that with the increase of noise level [see (a) and (b)], the ROC performance decreases, and the quantitative estimate AUC decreases from 0.99 in (a) to 0.9 in (b). The above two observations are consistent with our expectation. Panels (d) and (e) display the SIC value with respect to the tradeoff parameter λ that corresponds to the simulations of panels (a) and (b), respectively. The red circle indicates the best λ that gives the lowest SIC. By comparing (a) and (d) with (b) and (e), respectively, it can be seen that when the noise level is relatively low, the valley in the SIC curve tends to be flat [see panel (d)]. Therefore, the resulting λ is small, yielding high TPR [see panel (a)]. On the other hand, for high noise level, large λ [see panel (e)] will be yielded, delivering low FPR at the cost of TPR loss [see panel(b)]. Generally speaking, this simulation results demonstrate that the SIC can provide user with a reasonable λ at various noise level. Panel (f) displays a typical cost function with respect to the iteration number showing that the algorithm converges within 100 iterations.

2) Simulation II: In this simulation, we followed the NGS protocol to generate *hybrid* dataset that is more realistic.

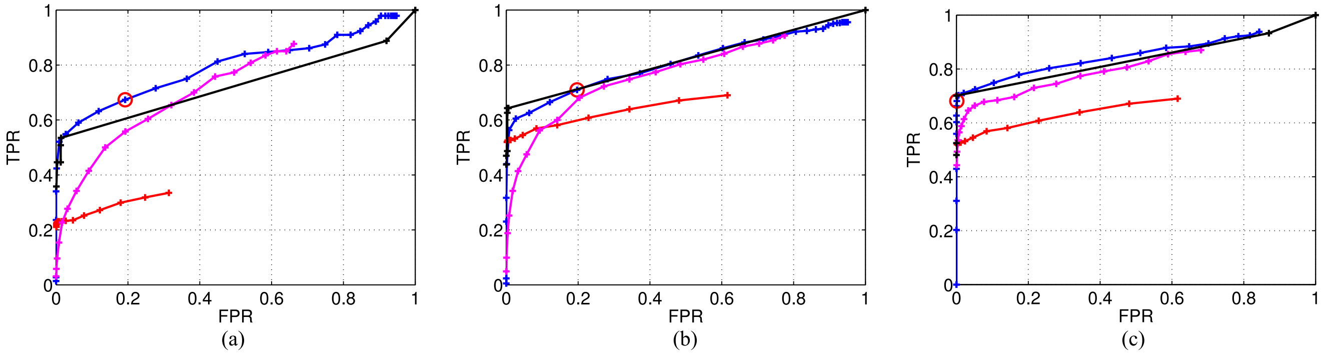


Fig. 3. ROCs of detecting CNVs in Simulation II. Blue, red, magenta, and black color represent the ROC of the proposed method, CoNIFER, XHMM, and cn.MOPS, respectively. The plus markers represent different levels of λ, *–threshold*, fourth/eighth parameter field of XHMM, and *–upperThreshold/lowerThreshold*. The red circle marks the one that yields lowest SIC. (a) Each CNV covers three to six exons, while in (b) and (c) each covers six. In (a) and (b), the total number of paired-end reads is 1e6, while in (c) this number is 1e7.

1) Download the *.FASTA* file of chromosome 21 of human reference genome hg19 and the corresponding *.bed* file that contains the loci of exons from UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgTables?org=human). Considering the distribution of human exon size, only exons with length larger than 100 bp were retained [51], yielding totally 1429 exon loci.
2) Call *RSVSim* [52] to simulate 60 samples of chromosome 21, and each contains a synthetic CNV located randomly, which is of length 1 kbp to 1 Mbp and covers three to six consecutive exon loci.
3) Call *wgsim* from the *samtools* package [53] to simulate 1 million short paired reads of length 70 bp for each sample chromosome, with base error rate 0.02, outer distance between the two ends 500 bp, standard deviation 50, mutation rate 0.001, indel fraction 0.15, and extended indel probability 0.3.
4) Call *bowtie* [54] to align the paired ends to the human reference genome hg19. Parallel alignment with four cores was enabled and the output was sorted, indexed, and converted to *.bam* format with *samtools*, yielding 60 bam files.
5) Call *rpkm* tool of CoNIFER [30] to calculate the RPKM (reads per thousand bases per million reads sequenced [55]) profile from each sample bam file. Since CoNIFER needs to import a *.bed* file that lists the loci of probes, we use the same one as in step 1).

Once RPKM profile of each sample chromosome was ready, the proposed methods, CoNIFER [30], XHMM [31], and cn.MOPS [21], were called to detect CNVs, respectively. To obtain a fair comparison for each method, the significant parameters which can greatly change the performance in terms of sensitivity and specificity were tuned to achieve the best performance, and the less significant parameters were left as default values. Since the low-frequency variations across the RPKM profiles are limited, parameter *–svd* of CoNIFER and *–numPCtoRemove* of XHMM were set to 3 such that the first three singular value decomposition (SVD) components were removed. Another crucial parameter *–threshold* of CoNIFER determines whether a probe harbors a CNV or not, so it plays a similar role as λ, *i.e.*, the decrease of this parameter will increase sensitivity, but at a loss of specificity at the same time. In our simulation, CoNIFER was called repeatedly with 15 distinct *–threshold* values following an arithmetic sequence from 0.1 to 1.5 with a common difference of 0.1. For XHMM, we found that the fourth and eighth fields in the parameter file, *i.e.*, the mean values of the Gaussian distributions corresponding to CNV deletion and duplication, respectively, function like the threshold and can control the tradeoff between sensitivity and specificity. So in our simulation study, the fourth field is set the negative value of the eighth field, which was increased from 0.36 to 0.3×1.2^{20} with a common ratio of 1.2, yielding the ROC curve of XHMM calls. For cn.MOPS, the parameters *–lowerThreshold* and *–upperThreshold* are the equivalents of XHMM, so we increased/decreased *–upperThreshold/lowerThreshold* from $2^{-2}/-2^{-2}$ to $2^{7}/-2^{7}$ with a common ratio of 2, yielding the ROC curve of cn.MOPS calls.

The results are displayed in Fig. 3. It is shown that *–threshold* of CoNIFER between 1.0 and 1.5 is a reasonable choice, and with further decrease of this value, the specificity decreases rapidly without much improvement of sensitivity. For XHMM and cn.MOPS, with the decrease of the upper threshold parameters, the sensitivity improves at the cost of the loss of specificity. There is a transient on the cn.MOPS ROC, where the FPR jumps from almost 0 to 0.92. Further experiments show that there is a critical value for *–upperThreshold* of cn.MOPS below which cn.MOPS achieves low specificity and *vice versa*. Panel (a) shows that the sensitivity of using CoNIFER is approximately half of the proposed method at the same specificity level. After more investigations, we found that this is because CoNIFER requires three or more consecutive probes that exceed the threshold to confirm a CNV. Considering that in Step 2 of this section, a synthetic CNV may cover three to six probes, in the case that one or more probes failed to exceed the threshold, a missing hit may happen. In addition, less probes a CNV covers, the higher chance it may be missed. Taking this factor into consideration, panel (b) shows the simulation result in which each CNV covers six probes. It is shown that the sensitivity of CoNIFER almost doubles at the same specificity level. In both panels (a) and (b), the λ chosen by the SIC is not satisfactory, which yields high FPR. This might be due to the relatively high noise level, since the sequencing coverage is 4.2. In panel (c) 10 million paired-end reads were sequenced; therefore, the overall coverage increased by a factor of 10. The result indicates that the performance (*i.e.*, ROC) of the proposed method increases slightly, but the estimation of λ improves greatly.

B. Real Data Study

We downloaded the whole-exome sequencing data of ten HapMap samples from the FTP of the 1000 Genomes Project (see [56], http://www.1000genomes.org/). The dataset includes four CEU samples: NA12287, NA12749, NA12776, and NA12828; and six YRI samples: NA19114, NA19129, NA19147, NA19190, NA19225, and NA19257. These samples have been studied previously [32], and CNV calls can be retrieved from the database of genomic variants (DGV, http://projects.tcag.ca/variation/). Raw sequencing reads were already aligned to human reference genome GRCh3h/hg19 with BWA [53] and stored in BAM files, so we downloaded these 10 BAM files and calculated RPKM values y for 186 741 exons from chromosomes 1 to 22. Among them, 45 038 probes were excluded from further analysis since their median RPKM is less than 1 [30].

The parameter p was first estimated, which is needed to run the proposed algorithm. We estimated p by fitting the observed distribution f_{obs} of residuals $\frac{y_{mn}}{\overline{y_{\cdot n}}} - \overline{y_{m\cdot}}$ with a theoretical one f_{theo}. The observed distribution was calculated via histogram of the data, while the theoretical one can be calculated according to (2).

Once f_{theo} and f_{obs} are known or estimated, their difference can be measured by utilizing the Kullback–Leibler (KL) divergence, which is defined as

$$D_{\mathrm{KL}}(f_{\mathrm{obs}} \| f_{\mathrm{theo}}) = \int_{-\infty}^{+\infty} \ln\left(\frac{f_{\mathrm{obs}}(x)}{f_{\mathrm{theo}}(x)}\right) f_{\mathrm{obs}}(x)\,\mathrm{d}x. \quad (21)$$

We explored p from 0.5 to 1.5 with a common difference of 0.01 [see Fig. 1(a)]. For each fixed p and corresponding $\sigma(p)$ estimated according to (3), f_{theo} can be calculated according to (2). Then the KL divergence of f_{theo} from f_{obs} was calculated according to (21). A low KL divergence value indicates a good fitting; therefore, the criterion of selecting p is to find the one that gives the lowest value. Fig. 1(a) shows that $p = 0.92$ yields the lowest KL divergence, while (b) and (c) show that the theoretical distribution with this p fits to the observed one extremely well.

Next, we run the proposed algorithm with estimate $p = 0.92$ to call candidate CNVs. The maximal iteration number was set to 200, and the penalty parameter λ was set to $\sigma^p \ln(MN)$ according to the analysis of SIC curve. To reduce false detection and fragmental region, CNV calls that cover less than three exons were filtered out, and then neighboring CNV calls with gap less than three exons were merged. At the end, the algorithm outputted 35 CNV calls in total. We also run CoNIFER, XHMM, and cn.MOPS to analyze the same dataset. For XHMM and CoNIFER, the first three SVD components were removed, and other parameters remain as default. CoNIFER, XHMM, and cn.MOPS outputted 49, 32, and 44 calls, respectively. Fromer [31] showed that there are two CNVs per individual on average, so the numbers of calls are within the reasonable range on the ten sample study.

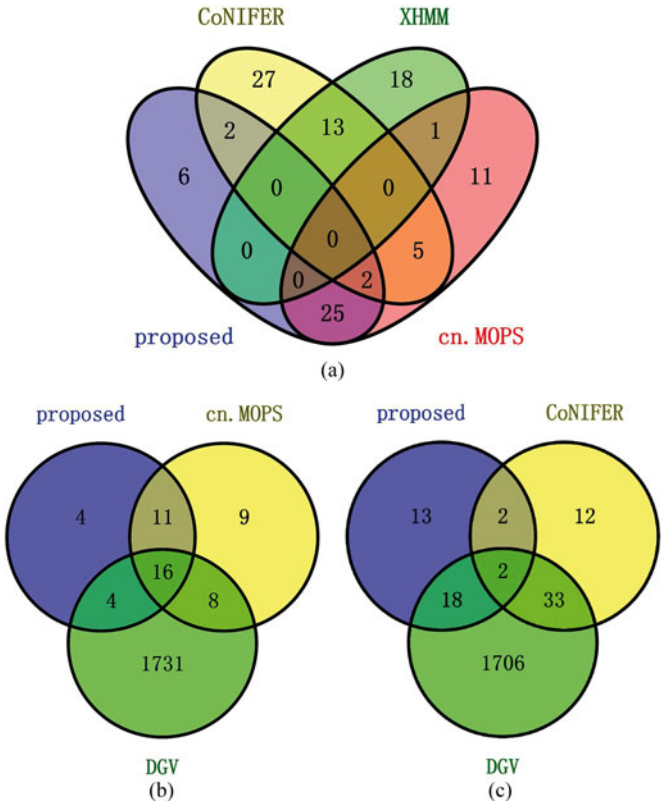


Fig. 4. Venn diagrams of CNV detection results among the proposed method, CoNIFER, XHMM, cn.MOPS, and DGV.

Finally, we compared the CNV calls of the proposed method with those of CoNIFER, XHMM, and cn.MOPS. The comparison result is displayed in the Venn diagram [57] of Fig. 4(a) and summarized in supplementary Table S1, which shows that out of 35 CNV calls, 29 (82.9%) overlap with other calls (CoNIFER, XHMM, or cn.MOPS). On the other hand, out of 49 CoNIFER calls, 32 XHMM calls and 44 cn.MOPS calls, 22 (44.9%) 14 (43.8%), and 33 (75.0%) overlap with the other calls, respectively. This shows that the proposed method achieves higher precision compared with other method, indicating more reliable detection performance. Note that there are two large overlaps: One is the intersection between the proposed method and cn.MOPS, and the other is between XHMM and CoNIFER. This observation indicates that the proposed method and cn.MOPS share similar performance, while CoNIFER and XHMM also share the similar performance. This is consistent with the simulation results as shown in Fig. 3, where blue and black ROCs have similar trajectory. It also supports the fact that XHMM and CoNIFER utilize PCA as the main approach to reduce baselines [31]. To further verify CNV calls, we compared the calls with DGV, and the overlaps are displayed in the Venn diagrams of Fig. 4(b) and (c). It is shown in (b) that among the 27 overlapping calls of using cn.MOPS and the proposed method, 16 (59.3%) calls are included in DGV. On the contrary, even though there are 53 DGV calls of the proposed method or CoNIFER [see Fig. (c)], only 2 are detected by both methods. This suggests that those methods are complementary in detecting CNVs. Note that since exome comprises 1% to 2% of the whole human genome [25], the majority of CNVs listed in DGV are not covered by exome sequencing. In the above comparison, only CNVs in DGV covered with probes were taken into account, including 1759 calls. Among those calls, only a small portion overlaps with our tested algorithms. The reason might be that DGV gathers all reported CNVs, which is a union set of several studies, so the retrieved calls might include some false detections. For real data, since there is a lack of ground truth, we use the F-scores [3] to qualitatively measure the detection result. The F-score is calculated as follows. If a CNV call has no overlap with DGV, the F-score is set to be 0; otherwise, $F = 2\frac{PR}{P+R}$, where P is the precision (the percent of the CNV call that overlaps with DGV) and R is the recall (the percent of DGV that overlaps with CNV calls). Note that F-score takes values between 0 (no overlapping at all) and 1 (perfect overlapping). Among the 20 overlapping CNVs of using the proposed method, 7 (35.0%) get F-score above 0.5, while for CoNIFER, XHMM, and cn.MOPS, this percentage is 34.3%, 25.0%, and 25.5%, respectively, indicating the good detection quality of the proposed method.

C. Discussion

The proposed method has the following three features. First, we assume that the ideal read depth of exome sequencing data can be approximated with a low-rank matrix, based on which a penalized matrix approximation model is proposed. CoNIFER actually used a similar assumption. In CoNIFER, the data matrix is first factorized with SVD, and then singular vectors with significant singular values are kept to approximate the raw data. Therefore, the approximation matrix used in CoNIFER is also low rank.

Second, real data analysis shows that the read depth of exome sequencing data can be fitted with the generalized Gaussian distribution, and this information can be incorporated into our model as a prior knowledge. The p parameter in the generalized Gaussian distribution affects the performance of the proposed method, which needs to be predefined before algorithm execution. To tune this parameter, we used the KL divergence to determine the optimal p value.

Third, it is worth to note that the exon size is not incorporated into our Model (1), since the read depth signal has taken this factor into account. Another factor that is not incorporated explicitly into the model is the G-C bias. A previous study showed that read depth signal is correlated with the G-C content [58], which is uneven along the genome. Therefore, the detection performance from single sample of whole genome sequencing degenerates if the G-C bias is not corrected. Since the model (1) dedicates to detect CNVs from a pool of samples, the effect of bias caused by the G-C distribution is reduced implicitly.

The convergence of the optimization algorithm is a critical issue in solving the proposed sparse model, which can impact computational time. There are two steps involved with the optimization: the estimation of $\boldsymbol{d}$ and $\boldsymbol{v}$. Daubechies *et al.* [43] reported that the IRLS converges to a local minimizer with superlinear convergence rate, and Nie *et al.* [59] provided a rigorous proof that IRLS for $\ell_{2,1}$-norm minimization converges to the global minimizer. So for a given $\boldsymbol{d}$, the IRLS estimate of $\boldsymbol{v}$ leads to the decrease of the cost function. For a given $\boldsymbol{v}$, an exhaustive search also leads to the decrease of the cost function. Since $\boldsymbol{d}$ and $\boldsymbol{v}$ are estimated alternatively and both estimates yield the decrease of the cost function, overall the algorithm converges to a local minimum. As demonstrated in our experiment [see Simulation I, Fig. 2(f)], the algorithm indeed converges very fast in practice. On a desktop with Intel i7 processor, the processing of the real dataset costs on average 5 s per iteration with 50 MB memory allocation.

IV. Conclusion

A novel method was proposed to detect CNVs from exome sequencing data. The method is based on the assumption that the read depth of sequencing data could be approximated with a low-rank sparse regression model. For the proposed model with a penalty term, an efficient numerical algorithm was applied based on the IRLS algorithm [43], [59]. The performance of the proposed method was tested and validated on both synthetic and real data. Our simulation studies demonstrate that the proposed method can achieve higher detection power than three published methods including CoNIFER, XHMM, and cn.MOPS, especially at the CNV regions where the number of available probes is limited. Experimental results on real data show that they can be well fitted with the proposed model. In addition, the outputted CNVs detected with the proposed method have higher overlapping percentage compared with those from CoNIFER, XHMM, and cn.MOPS calls, showing that it can detect CNV with higher precision.

In order to make our developed algorithms and tools to be publicly available so other researchers can use and compare with their methods, we publish the source codes at http://www.tulane.edu/ wyp/software/Exon_CNV.m

References

[1] J. Korbel *et al.*, “Paired-end mapping reveals extensive structural variation in the human genome,” *Science*, vol. 318, pp. 420–426, Oct. 2007.

[2] J. Duan *et al.*, “Comparative studies of copy number variation detection methods for next generation sequencing technologies,” *Plos One*, vol. 8, no. 3, pp. 1–12, 2013.

[3] P. Medvedev *et al.*, “Computational methods for discovering structural variation with next-generation sequencing,” *Nature Methods*, vol. 6, pp. S13–S20, Nov. 2009.

[4] A. Magi *et al.*, “Bioinformatics for next generation sequencing data,” *Genes*, vol. 1, pp. 294–307, 2010.

[5] A. Rovelet-Lecrux *et al.*, “APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy,” *Nature Genet.*, vol. 38, no. 1, pp. 24–26, Jan. 2006.

[6] J. Sebat *et al.*, “Strong association of de novo copy number mutations with autism,” *Science*, vol. 316, pp. 445–449, Apr. 2007.

[7] P. Campbell *et al.*, “Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing,” *Nature Genet.*, vol. 40, pp. 722–729, Jun. 2008.

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

[9] T.-L. Yang *et al.*, “Genome-wide copy-number-variation study identified a susceptibility gene, UGT2B17, for osteoporosis,” *Am. J. Hum. Genet.*, vol. 83, no. 6, pp. 663–674, Dec. 2008.

[10] J. Freeman *et al.*, “Copy number variation: New insights in genome diversity,” *Genome Res.*, vol. 16, pp. 949–961, Aug. 2006.

[11] P. Stankiewicz and J. R. Lupski, "Structural variation in the human genome and its role in disease," *Annu. Rev. Med.*, vol. 61, pp. 437–455, 2010.

[12] E. Schrock *et al.*, "Multicolor spectral karyotyping of human chromosomes," *Science*, vol. 273, no. 5274, pp. 494–497, 1996.

[13] M. Speicher *et al.*, "Karyotyping human chromosomes by combinatorial multi-fluor fish," *Nature Genet.*, vol. 12, no. 4, pp. 368–375, 1996.

[14] J. Chen and Y.-P. Wang, "A statistical change point model approach for the detection of DNA copy number variations in array CGH data," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 6, no. 4, pp. 529–541, Oct. 2009.

[15] D. Chiang *et al.*, "High-resolution mapping of copy-number alterations with massively parallel sequencing," *Nature Methods*, vol. 6, pp. 99–103, Jan. 2009.

[16] C. Xie and M. T. Tammi, "CNV-seq, a new method to detect copy number variation using high-throughput sequencing," *BMC Bioinformat.*, vol. 10, pp. 1–9, 2009.

[17] S. Yoon *et al.*, "Sensitive and accurate detection of copy number variants using read depth of coverage," *Genome Res.*, vol. 19, pp. 1586–1592, Sep. 2009.

[18] V. Boeva *et al.*, "Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization," *Bioinformatics*, vol. 27, no. 2, pp. 268–269, Jan. 2011.

[19] C. Miller *et al.*, "ReadDepth: A parallel R package for detecting copy number alterations from short sequencing reads," *PLoS ONE*, vol. 6, pp. 1–7, 2011.

[20] A. Abyzov *et al.*, "CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing," *Genome Res.*, vol. 21, no. 6, pp. 974–984, Jun. 2011.

[21] G. Klambauer *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.*, vol. 40, no. 9, pp. 1–14, 2012.

[22] J. Duan *et al.*, "CNV-TV: A robust method to discover copy number variation from short sequencing reads," *BMC Bioinformat.*, vol. 14, no. 150, pp. 1–12, 2013.

[23] K. Wetterstrand. (2015). DNA sequencing costs: Data from the NHGRI genome sequencing program (GSP). [Online]. Available: www.genome.gov/sequencingcosts

[24] E. Karakoc *et al.*, "Detection of structural variants and indels within exome data," *Nature*, vol. 9, no. 2, pp. 176–178, 2012.

[25] J. A. Tennessen *et al.*, "Evolution and functional impact of rare coding variation from deep sequencing of human exomes," *Science*, vol. 337, no. 64, pp. 64–69, 2012.

[26] J. Sathirapongsasuti *et al.*, "Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV," *Bioinformatics*, vol. 27, no. 19, pp. 2648–2654, 2011.

[27] A. Alkodsi *et al.*, "Comparative analysis of methods for identifying somatic copy number alterations from deep sequencing data," *Briefings Bioinformat.*, vol. 16, pp. 242–254, 2015.

[28] V. Plagnol *et al.*, "A robust model for read count data in exome sequencing experiments and implications for copy number variant calling," *Bioinformatics*, vol. 28, no. 21, pp. 2747–2754, 2012.

[29] M. I. Love *et al.*, "Modeling read counts for CNV detection in exome sequencing data," *Stat. Appl. Genet. Mol. Biol.*, vol. 10, pp. 1–28, 2011.

[30] N. Krumm *et al.*, "Copy number variation detection and genotyping from exome sequence data," *Genome Res.*, vol. 22, pp. 1525–1532, 2012.

[31] M. Fromer *et al.*, "Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth," *Am. J. Human Genet.*, vol. 91, pp. 597–607, 2012.

[32] D. F. Conrad *et al.*, "Origins and functional impact of copy number variation in the human genome," *Nature*, vol. 464, no. 7289, pp. 704–712, 2009.

[33] R. Chartrand, "Nonconvex splitting for regularized low-rank + sparse decomposition," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5810–5819, Nov. 2012.

[34] J. Idier, Ed., *Bayesian Approach to Inverse Problems*. New York, NY, USA: Wiley, Apr. 2008.

[35] A. J. Iafrate *et al.*, "Detection of large-scale variation in the human genome," *Nature Genet.*, vol. 36, no. 9, pp. 949–951, Sep. 2004.

[36] M. Nikolova, "Local strong homogeneity of a regularized estimator," *SIAM J. Appl. Math.*, vol. 61, no. 2, pp. 633–658, 2000.

[37] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, Apr. 1995.

[38] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. R. Statist. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.

[39] B. Efron *et al.*, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.

[40] D. M. Malioutov *et al.*, "Homotopy continuation for sparse signal representation," in *Proc. IEEE Int. Accoust. Speech, Signal Process.*, Philadephia, PA, USA, Mar. 2005, vol. V, pp. 733–736.

[41] D. L. Donoho and Y. Tsaig, "Fast solution of l_1-norm minimization problems when the solution may be sparse," *IEEE Trans. Inf. Theory*, vol. 54, no. 11, pp. 4789–4812, Nov. 2008.

[42] C.-H. Zhang, "Discussion: One-step sparse estimates in nonconcave penalized likelihood models," *Ann. Statist.*, vol. 36, no. 4, pp. 1509–1533, 2008.

[43] I. Daubechies *et al.*, "Iteratively reweighted least squares minimization for sparse recovery," *Commun. Pure Appl. Math.*, vol. 63, no. 1, pp. 1–38, 2010.

[44] X. Chen *et al.*, "Smoothing proximal gradient method for general structured sparse regression," *Ann. Appl. Stat.*, vol. 6, no. 2, pp. 719–752, 2012.

[45] S. S. Chen *et al.*, "Atomic decomposition by basis pursuit," *SIAM Rev.*, vol. 43, no. 1, pp. 129–159, 2001.

[46] P. Hansen, "Analysis of discrete ill-posed problems by means of the L-curve," *SIAM Rev.*, vol. 34, pp. 561–580, 1992.

[47] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Control*, vol. AC-19, no. 6, pp. 716–723, Dec. 1974.

[48] G. Schwarz, "Estimating the Dimension of a Model," *Ann. Stat.*, vol. 6, pp. 461–464, 1978.

[49] K. E. Markon and R. F. Krueger, "An empirical comparison of information-theoretic selection criteria for multivariate behavior genetic models," *Behavior Genet.*, vol. 34, no. 6, pp. 593–610, 2004.

[50] R. Xi *et al.*, "Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion," *Proc. Nat. Acad. Sci.*, vol. 108, no. 46, pp. E1128–E1136, Nov. 2011.

[51] International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860–921, 2001.

[52] C. Bartenhagen and M. Dugas, "RSVSim: An R/Bioconductor package for the simulation of structural variations," *Bioinformatics*, vol. 29, no. 13, pp. 1679–1681, 2013.

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

[54] B. Langmead *et al.*, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biol.*, vol. 10, no. 3, pp. 1–10, 2009.

[55] A. Mortazavi *et al.*, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.

[56] The 1000 Genomes Project Consortium, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–1073, Oct. 2010.

[57] J. C. Oliveros. (2015). Venny. An interactive tool for comparing lists with Venn's diagrams. [Online]. Available: http://bioinfogp.cnb.csic.es/tools/venny/index.html

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

[59] F. Nie *et al.*, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," in *Adv. Neural Inform. Process. Syst.*, 2010, pp. 1–9.

Junbo Duan (M'15) received the B.S. degree in information engineering and the M.S. degree in communication and information system from Xi'an Jiaotong University, Xi'an, China, in 2004 and 2007, respectively, and the Ph.D. degree in signal processing from Université Henry Poincaré, Nancy, France, in 2010.

After graduation, he was a Postdoctoral Fellow with the Department of Biomedical Engineering and Biostatistics & Bioinformatics at Tulane University, USA, until 2013. He is currently an Assistant Professor with the Department of Biomedical Engineering, Xi'an Jiaotong University. His major research interests include probabilistic approaches to inverse problems in biomedical engineering and bioinformatics.

Mingxi Wan (M'01) was born in Hubei, China, in 1962. He received the B.S. degree in geophysical prospecting from Jianghan Petroleum Institute, Jingzhou, China, in 1982, and the M.S. and Ph.D. degrees in biomedical engineering from Xi'an Jiaotong University, Xi'an, China, in 1985 and 1989, respectively.

He is currently a full Professor with the Department of Biomedical Engineering, Xi'an Jiaotong University. From 1995 to 1996, he was a Visiting Scholar and Adjunct Professor at Drexel University, Philadelphia, PA, USA, and Pennsylvania State University, University Park, PA. From 2000 to 2001, he was a Visiting Scholar with the Department of Biomedical Engineering, University of California, Davis, CA, USA. From 2000 to 2010, he was the Dean of the School of Life Science and Technology, Xi'an Jiaotong University. He is an Author and Coauthor of more than 100 peer-reviewed publications in international journals and five books about medical ultrasound. His current research interests include voice science, ultrasonic imaging, especially in tissue elasticity imaging, contrast and tissue perfusion evaluation, therapeutic ultrasound, and theranostics.

Dr.Wan has received several important awards from the Chinese government and university.

Hong-Wen Deng received the Bachelor's degree in ecology and environmental biology and the Master's degree in ecology and entomology from Peking University, Beijing, China. He received the Master's degree in mathematical statistics and the Ph.D. degree in quantitative genetics from the University of Oregon, Eugene, OR, USA.

He was a Postdoctoral Fellow at the Human Genetics Center, University of Texas, Houston, TX, USA, where he conducted postdoctoral research in molecular and statistical population/quantitative genetics. He received a Hughes Fellowship at the Institute of Molecular Biology, University of Oregon. He was a Professor of medicine and biomedical sciences at Creighton University Medical Center, a Professor of orthopedic surgery and basic medical science, and the Franklin D. Dickson/Missouri Endowed Chair in orthopedic surgery at the School of Medicine, University of Missouri-Kansas City. He is currently the Chair of Tulane Biostatistics and Bioinformatics Department and the Director of the Center of Bioinformatics and Genomics. He widely published with nearly 500 peer-reviewed articles, ten book chapters, and three books. His research interest includes biostatistics, bioinformatics, genomics and epigenomics of osteoporosis, and obesity.

Dr. Deng received multiple NIH R01 awards and an NIH P50 award, and multiple honors for his research.

Yu-Ping Wang (SM'06) received the B.S. degree in applied mathematics from Tianjin University, Tianjin, China, in 1990, and the M.S. degree in computational mathematics and the Ph.D. degree in communications and electronic systems from Xi'an Jiaotong University, Xi'an, China, in 1993 and 1996, respectively.

After graduation, he had visiting positions at the Center for Wavelets, Approximation and Information Processing of the National University of Singapore and Washington University Medical School in St. Louis. From 2000 to 2003, he was a Senior Research Engineer at Perceptive Scientific Instruments, Inc., and then Advanced Digital Imaging Research, LLC, Houston, TX, USA. 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 the Tulane University School of Science and Engineering, New Orleans, LA, USA, and the School of Public Health and Tropical Medicine, New Orleans. He is also a Member of the Tulane Center of Bioinformatics and Genomics, Tulane Cancer Center and Tulane Neuroscience Program. He has served on numerous program committees and NSF/NIH review panels, and served as Editors for several journals such as *Neuroscience Methods.* His research interests include computer vision, signal processing, and machine learning with applications to biomedical imaging and bioinformatics, where he has about 150 peer-reviewed publications. .