Zhang *et al. BMC Bioinformatics* (2016) 17:384
DOI 10.1186/s12859-016-1239-7

BMC Bioinformatics

RESEARCH ARTICLE Open Access

nbCNV: a multi-constrained optimization model for discovering copy number variants in single-cell sequencing data

Changsheng Zhang, Hongmin Cai* , Jingying Huang and Yan Song

Abstract

Background: Variations in DNA copy number have an important contribution to the development of several diseases, including autism, schizophrenia and cancer. Single-cell sequencing technology allows the dissection of genomic heterogeneity at the single-cell level, thereby providing important evolutionary information about cancer cells. In contrast to traditional bulk sequencing, single-cell sequencing requires the amplification of the whole genome of a single cell to accumulate enough samples for sequencing. However, the amplification process inevitably introduces amplification bias, resulting in an over-dispersing portion of the sequencing data. Recent study has manifested that the over-dispersed portion of the single-cell sequencing data could be well modelled by negative binomial distributions.

Results: We developed a read-depth based method, *nbCNV* to detect the copy number variants (CNVs). The nbCNV method uses two constraints-sparsity and smoothness to fit the CNV patterns under the assumption that the read signals are negatively binomially distributed. The problem of CNV detection was formulated as a quadratic optimization problem, and was solved by an efficient numerical solution based on the classical alternating direction minimization method.

Conclusions: Extensive experiments to compare nbCNV with existing benchmark models were conducted on both simulated data and empirical single-cell sequencing data. The results of those experiments demonstrate that nbCNV achieves superior performance and high robustness for the detection of CNVs in single-cell sequencing data.

Keywords: Copy number variants, Read depth, Negative binomial distribution, Sparsity, Smoothness, ADMM

Background

Copy number variants (CNVs), which constitute a major form of DNA structural variation, have been shown to be closely related to several diseases, including autism [11], schizophrenia [29] and cancer [5, 8, 15, 21]. Comparative genomic hybridization and fluorescence in situ hybridization have been used to detect CNVs of particular genes or fragments [26] but are limited in terms of resolution. To profile genome-wide copy number (CN) landscapes, these techniques have consequently been replaced by next-generation sequencing (NGS) technologies [9]. Because it uses bulk DNA from tissue samples, however, traditional sequencing provides an average signal from millions of cells and is thus of limited utility for the characterization of tumor heterogeneity at the single-cell level.

An innovative technique, single-cell sequencing (SCS), was developed to address key issues in cancer studies, including measurement of mutation rates, tracing of cell lineages, resolution of intra-tumor heterogeneity and elucidation of tumor evolution [21, 22]. SCS combines flow sorting of single cells, whole-genome amplification (WGA) and NGS to characterize the genome-wide CN of single cells. Existing WGA techniques, such as degenerate oligonucleotide primed-polymerase chain reaction [30], multiple displacement amplification [17] and multiple annealing looping-based amplification cycling [36], inevitably introduce amplification bias to varying degrees when the whole genome of a single cell is amplified to microgram levels for NGS [13, 28]. Such technical

*Correspondence: hmcai@scut.edu.cn
School of Computer Science & Engineering, South China University of Technology, 510006 Guangzhou, China

noise due to amplification bias is over-dispersed and is different from that of bulk sequencing, which does not involve amplification. There are two main strategies that use NGS data to detect CNVs: read depth (RD)-based and read pair (RP)-based methods [20]. To the best of our knowledge, RD-based methods are arguably popular for CNV detection. Furthermore, CNV detection using SCS data requires only sparse sequence coverage to economically accommodate numerous single cells [4].

The analysis pipeline of RD-based methods consists of data preparation (optional), data normalization (optional), CNV region identification (core) and CN profile estimation (optional) [19]. Briefly speaking, the reference genome is divided into equally or variably sized, non-overlapping bins for computing read counts (RCs) in each bin along the whole genome. The RD in each bin is generated by the corresponding RC divided by the average RC for the whole genome. The RD signal is then normalized using strategies such as lowness smoothing based on guanine-cytosine (GC) content. Different segmentation algorithms are used to detect the CNV regions. After their detection, the CNV regions can be translated into a CN profile using available ploidy information or by other methods [3, 19]. Several existing RD-based benchmark CNV detection methods, which we later apply for comparison, are described as follows.

DNAcopy [27] implements a classical circular binary segmentation (CBS) [25] algorithm to segment RD data and identifies abnormal genomic regions. The basic idea of CBS is to translate a noisy-intensity RD signal into regions of equal CNs followed by binary segmentation. Copynumber [24] is a highly efficient algorithm that offers a unified framework to segment RD data from single or multiple samples. This approach combines least squares principles with a suitable penalization scheme for a given number of breakpoints to detect CN profiles. The above two methods do not require data preparation and normalization. Control-free copy number and allelic content caller (Control-FREEC) [6] is a systematic CNV detection package consisting of data preparation, normalization, CNV region identification, and profile estimation. Control-FREEC segments the whole reference genome into equally sized, non-overlapping bins. It then computes the RD of the tested sample in each bin. If a control sample is not supplied, Control-FREEC uses the GC content in each bin to achieve data normalization. For CNV detection, Control-FREEC uses least absolute shrinkage and selection operator (LASSO) regression. CNVnator [1] also divides the whole reference genome (hg18 or hg19) into continuous, non-overlapping equal-sized bins. Normalization is achieved by averaging the RD signal over each bin with respect to GC content, and CNV region identification is based on mean-shift tracking.

Most of the earlier analysis in sequencing analysis assumed that the RDs are following Gaussian distribution [2, 24, 31] or Poisson distribution [1, 10, 14]. However, experimental analysis in the absolute numbers of mRNA molecules by single-cell sequencing manifested [12] that the counts could be accurately characterized by negative binomial distribution. To this aim, we firstly generated an illustrative example to characterize the statistical distribution of real RDs by single-cell sequencing technology. Real sequencing data from a normal cell (accession number SRR052047) [21] were preprocessed to obtain RC data with 50000 variable bins. The frequency histogram of the RC data is shown in Fig. 1(a). The distribution was approximated by the Poisson, Gaussian and negative binomial probability density functions through maximum likelihood estimation (MLE). The estimated mean value of the Poisson distribution was $\lambda = 62$. For the negative binomial distribution, the estimated mean value and dispersion coefficient are $\mu = 61.94$ and $\alpha = 9.544$, respectively. For Gaussian distribution, the estimated mean value was $\mu = 61.950$ and the estimated standard deviation $\sigma = 20.703$. This figure clearly demonstrated that the frequency histogram of the real RC data could be nicely characterized by a negative binomial distribution. For further comparison, we also measured the overlap ratio between the real RC and its approximations with different bin counts, ranging from 10000 to 210000. The overlap ratio was calculated as,

$$f(x) = \frac{A(x) \bigcap B(x)}{A(x) \bigcup B(x)} \tag{1}$$

where A denotes the frequency of empirical RC x and B is the approximated probability density function. The result was shown in Fig. 1(b). The ratio between the empirical RC and the one approximated by NB distribution was dramatically higher than by the Poisson distribution. When the bin count was larger than 50000, the ratio is more than 0.9 by the NB distribution. In comparison, the ratio was low to 0.4 by Poisson distribution. For Gaussian distribution, the ratio was rising when the bin count ranged from 10000 to 30000 and the ratio under 30000 was as higher as the one by NB distribution. However, it began dropping continuously with the bin count rising continuously. Therefore, the RC distribution could be accurately characterized by the negative binomial distribution.

To this end, a novel model called *nbCNV* was proposed in this paper to detect CNVs using SCS data. The nbCNV model uses negative binomial distributions to approximate loci along the whole genome. We incorporate two

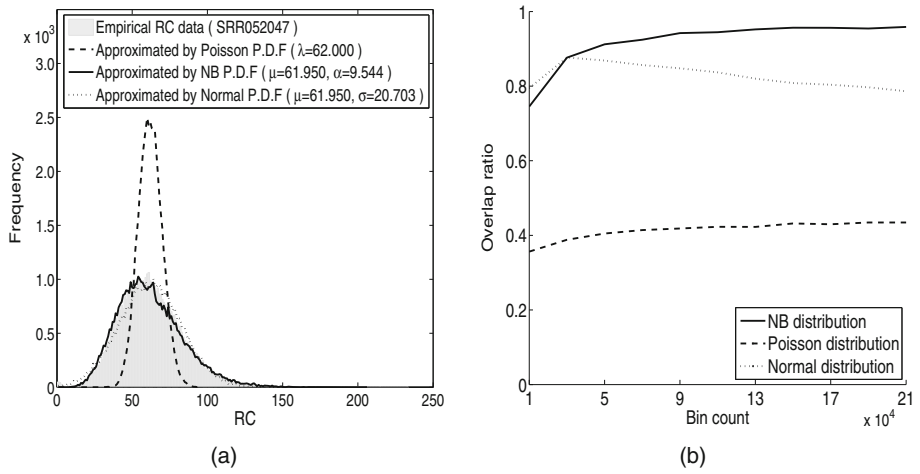


Fig. 1 Modeling of single-cell sequencing RC data and selection of a suitable bin count for experiments. **a** Frequency histogram of RC data (accession number SRR052047), shown in light green. For comparison, the frequencies of simulated RCs generated by Poisson , Gaussian and negative binomial distributions are indicated by brown, yellow or dark green lines, respectively. The RCs of the SCS data can be accurately characterized by a negative binomial distribution. The estimated parameters by MLE for each P.D.F are shown in parenthesis. **b** Relationship between bin count (horizontal axis) and overlap ratio (vertical axis). The empirical sequencing data could be better approximated by a negative binomial distribution. Such nice approximation was insensitive to the bin count. Therefore, the proposed nbCNV is hoping to be applied in various bin count setting

constraints of sparsity and smoothness to fit the CNV patterns. The CNV detection problem is then formulated by a quadratic optimization model. The proposed nbCNV uses an efficient numerical scheme based on the classical alternating direction minimization method (ADMM) to achieve efficiency. Since SCS data analysis requires carefully data preprocessing, a recently published SCS protocol [3] was modified to fit with the proposed nbCNV detection method. We have built a systematical pipeline for single-cell sequencing analysis. Considering the inherent contradiction between the quality [7] and resolution of CNVs detected with RD-based methods, nbCNV can adaptively select the most suitable total number of bins according to user preference. Once ploidy information is provided, the CNV regions detected by nbCNV can be translated into a CN profile.

The rest of this article is organized as follows. The underling mathematical models of nbCNV and its numerical solution are described in section "Methods". We then evaluate and demonstrate the efficiency of nbCNV compared with several benchmark methods using both simulated and real SCS datasets in section "Results and discussion". Finally, we conclude the paper in section "Conclusions".

Methods

Data preprocessing

To achieve data preparation and normalization, we modified a previously reported protocol for genome-wide CN analysis of single cells [3]. The steps are briefly summarized here. We first downloaded a sequencing file from the National Center for Biotechnology Information (NCBI) short read archive (SRA) and used the bowtie2 alignment tool [16] to map the millions of short reads to the GRCh37 human reference genome. Bins of variable sizes were used to segment the whole genome. Bin boundaries were decided by the length of reads used in the CNV analysis. For example, simulated reads of length k are generated along the whole genome one base pair at a time when the length of reads is k bp. Thus, the total number of simulated reads is $L - k + 1$, where L is the length of a chromosome. In our experiments, nearly three billion simulated reads were aligned back to the reference genome and all unique mapping read positions were retained. To divide the whole genome into variable-sized segments, each segment except the last one in each chromosome was possessed to have the same number of uniquely mappable positions. To achieve uniform bins, parameters for bowtie2 in each run were set to be equal. The RD signal was further normalized by locally-weighted polynomial regression (using LOWESS smoother, a function in the R language) and linear interpolation based on the GC content in each bin [3].

One of the important issues rising in above procedures is to decide the size of bin count. To investigate its dynamic relationship with the quality of the RC data, a quantitative measurement known as the multiple absolute pairwise difference (MAPD) threshold [7] was used to quantify the data quality. The MAPD is defined as $median(\left|\log_2 x_{i+1} - \log_2 x_i\right|)$, where x_i denotes the RC

signal at i-th position. A larger MAPD value implies of lower quality of the real RC data and less credibility of the following CNV detection. The relationship between data quality and bin count is shown in Fig. 2. As is evident from the figure, the quality of the RC data drops quickly when the bin count increases. It is due to the RC data tends to be more dispersed (of lower quality) if being preprocessed under a larger bin count.

In analyzing read depth data, a larger bin count is admired to achieve higher resolution. Provided with enough sequencing depth, high resolution analysis makes it accurate in CNV detection. However, if the sequencing depth is low, a larger bin count will deteriorate the data quality and make the data analysis less reliable. In order to find the accurate value in balancing the high resolution versus good quality as well as large overlap ratio $f(x)$, a simple maximization scheme is defined as following:

$$\max_x \left\{(1-\alpha)g(x) - \alpha k(x)\right\}$$

where $g(\cdot)$ is a polynomial functions aiming to fitting the *MPAD* function. $k(\cdot)$ is the overlap ratio defined in Eq. (1). The parameter of α is a trade-off parameter, ranging from 0.20 to 0.30. In our experiments, α was set to be 0.208 and it is corresponding to a bin count of 50000. When the bin count is 50000, the ratio between the empirical RC and the one approximated by NB distribution is higher than 90% and the MAPD value is closer to the maximum tolerable value (0.45). Therefore, the bin count of 50000 was selected to achieve a nice balance between the data quality and resolution of detection.

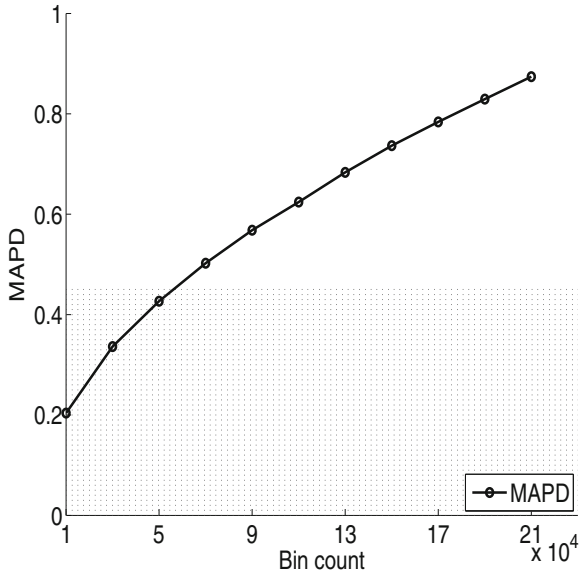


Fig. 2 Relationship between bin count (horizontal axis) and MAPD values (vertical axis). With increasing bin count, the quality of the data drops rapidly. A suitable bin count should thus be carefully selected by balancing the high resolution versus good quality as well as large overlap ratio. In our experiments, the maximum tolerable MAPD was set at 0.45 (stippled area) and the bin count was 50000

Problem formulation

Mathematically, let $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ be the observed RD signal in a bin, with $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ representing the corresponding reconstructed CN. We wish to determine the CN $\boldsymbol{x}$ that is most likely given by RD $\boldsymbol{y}$. By Bayes's Law:

$$P(\boldsymbol{x}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{x})P(\boldsymbol{x})}{P(\boldsymbol{y})}.$$

estimation of the CN $\boldsymbol{x}$ could be derived by maximizing the posterior probability $P(\boldsymbol{y}|\boldsymbol{x})P(\boldsymbol{x})$. Assuming a negative binomial distribution at each genome position t with mean parameter x and over-dispersed parameter α, we have:

$$P(y_t) = \frac{\Gamma(y_t+\alpha)}{y_t!\,\Gamma(\alpha)}\left(\frac{x_t}{x_t+\alpha}\right)^{y_t}\left(\frac{\alpha}{x_t+\alpha}\right)^{\alpha},$$

where α is an over-dispersed parameter that must be estimated empirically. For ease of model derivation, we temporarily assume that its value is known a priori and elaborate its estimation later.

If we assume that the values of y at position t are independent, then:

$$P(\boldsymbol{y}|\boldsymbol{x}) = \prod_t \frac{\Gamma(y_t+\alpha)}{y_t!\,\Gamma(\alpha)}\left(\frac{x_t}{x_t+\alpha}\right)^{y_t}\left(\frac{\alpha}{x_t+\alpha}\right)^{\alpha}.$$

Considering the characteristics of CNVs, we further require that the prior distribution on CN (after standardization by subtracting its mean value) satisfies assumptions related to two characteristics:

Smoothness: CNs at contiguous chromosome positions are similar except for abrupt changes between different segments;

Sparsity: CN variants are less common than invariants.

Mathematically, the above two characteristics can be penalized by:

$$P(\boldsymbol{x}) = \exp\left\{-\int(\lambda_1|\nabla\boldsymbol{x}| + \lambda_2|\boldsymbol{x}|)d\Omega\right\},$$

where λ_1 and λ_2 are trade-off parameters for respectively controlling the sparsity and smoothness of the CN function. The integration operation takes value along the genome on each bin Ω.

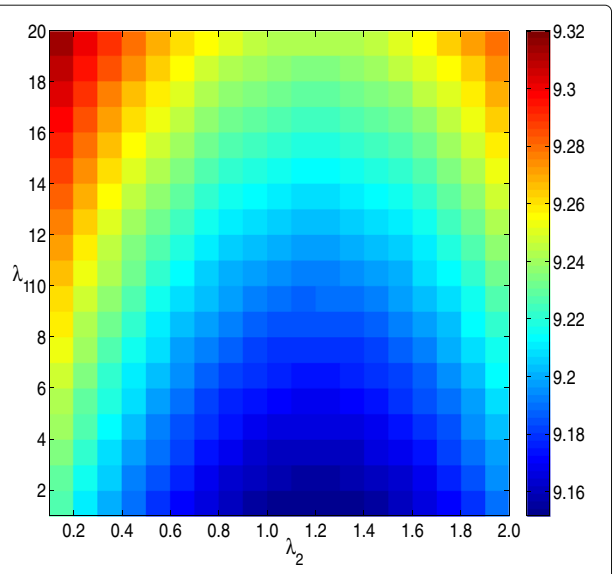


Fig. 3 Heatmap of Euclidean distance between the fitter signals and the real copy number signals with respect to the two parameters used by nbCNV. The horizontal axis stands for λ_1 and the vertical axis stands for λ_2. A smaller Euclidean distance implies a better detection performance

Finally, we minimize $-\log(P(\boldsymbol{y}|\boldsymbol{x})P(\boldsymbol{x}))$ to seek the maximum posterior probability on $\boldsymbol{x}$:

$$\min_{\boldsymbol{x},\alpha} \quad \{-\log\Gamma(\boldsymbol{y}+\alpha)+\log\Gamma(\alpha)-\alpha\log(\alpha) \qquad (2)$$
$$+\ (\boldsymbol{y}+\alpha)\log(\boldsymbol{x}+\alpha)^{+}-\boldsymbol{y}\log x^{+}+\lambda_1|\nabla\boldsymbol{x}|+\lambda_2|\boldsymbol{x}|\},$$

where $x^{+} = \max\{0,x\}$. However, minimization of this problem to have optimal x and α is infeasible because of the presence of the hyperbolic function Γ. If one uses the gradient descent method, the computation time needed to approximate the optimal solution will be very large. To alleviate this problem, we use a simple MLE-based method to estimate the value of α; thus, Eq. (2) can be simplified as

$$\min_{x}\left\{\sum_{t}\{(y_t+\alpha)\log(x_t+\alpha)^{+}-y_t\log x\}^{+}\right.$$
$$\left.+\int(\lambda_1|\nabla x|+\lambda_2|x|)d\Omega\right\}, \qquad (3)$$

where $x^{+} = \max\{0,x\}$. Once we have obtained the legitimate CN signal, its variants can be easily derived using simple thresholds.

Numerical solution

The minimization problem (3) is actually a quadratic optimization constrained both by a total variational norm and a l_1 norm. Such minimization problems are widely encountered in various areas, including signal processing and image recovery [23]. Because the optimization problem (3) is convex, multiple standard optimization methods are available for its solution, such as majority minimization [33, 34] and the Lasso approach [11, 35]. Because of the high volume of the sequencing data, however, an efficient numerical solution is desirable for practical usage. This paper proposes to solve Eq. (3) within the framework of the Alternating Direction Method of Multipliers (ADMM) method [23, 35]. The most attractive characteristic of ADMM is its ability to decompose a complex problem into favorably separable subproblems that can then be efficiently solved individually.

Let $g_1(\boldsymbol{x}) = (\boldsymbol{y}+\alpha)\log(\boldsymbol{x}+\alpha)-\boldsymbol{y}\log\boldsymbol{x}$, $g_2 = \mathbf{1}_{+}(\cdot)$, $g_3 = \lambda_1\|\cdot\|_1$, $g_4 = \lambda_2\|\cdot-c\|_1$, where $\mathbf{1}_{+}$ is the indicator function for positive real numbers:

$$\mathbf{1}_{+}(x) = \begin{cases} 0, & x > 0 \\ +\infty, & \text{otherwise.} \end{cases}$$

Let $\boldsymbol{G} = [\boldsymbol{I};\boldsymbol{I};\nabla;\boldsymbol{I}]^T$, with $\boldsymbol{I}$ being the identity matrix and ∇ the usual difference matrix. The minimization Eq. (3) can then be accordingly rewritten as:

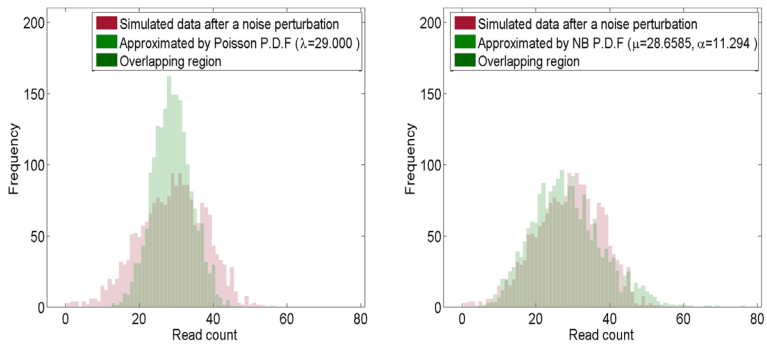


Fig. 4 Frequency histograms of contaminated RCs for the SRR052047 in chromosome-21 with implanted CNV sequences. The distribution of contaminated RCs can be better fitted by a negative binomial distribution than a Poisson distribution and is closer to that of empirical data

Table 1 Quantitative evaluation of the four tested methods in 100 simulation datasets. The best performance was highlighted in bold

Methods	Measurements			
	Accuracy	Precision	Sensitivity	Specificity
CNVnator	85.55 ± 6.56 %	88.66 ± 6.69 %	74.98 ± 15.96 %	92.24 ± 5.39 %
nbCNV	**91.83** ± **0.98** %	**94.85** ± **1.05** %	**84.78** ± **2.17** %	96.77 ± 0.69 %
control-FREEC	87.59 ± 2.10 %	94.43 ± 1.13 %	73.32 ± 6.15 %	**96.88** ± **0.93** %
poiCNV	89.41 ± 1.77 %	92.76 ± 0.89 %	80.61 ± 4.89 %	95.58 ± 0.72 %

$$\begin{aligned}\mathbb{L}(\boldsymbol{x}) &= \min_{\boldsymbol{x}} \sum_{t}^{m} \big\{(y_t+\alpha)\log(x_t+\alpha)^+ - y_t \log x_t^+ \\ &\quad + \mathbf{1}_+(x_t)) + \lambda_1\|\nabla x\|_1 \\ &\quad + \lambda_2\|x_t - c\|_1\big\} \\ &= \min_{\boldsymbol{x}} f_1(\boldsymbol{x}) + f_2(\boldsymbol{G}\boldsymbol{x}),\end{aligned}$$

where $f_1(\boldsymbol{x}) = 0, f_2(\boldsymbol{x}) = \sum_{j=1}^{4} g_j(\boldsymbol{x})$. After introducing a slack variable $\boldsymbol{u} = \boldsymbol{G}\boldsymbol{x}$, the augmented Lagrange function for $\mathbb{L}(\boldsymbol{x})$ is:

$$\mathbb{L} = \min_{\boldsymbol{x}} f_1(\boldsymbol{x}) + f_2(\boldsymbol{u}) + \frac{\boldsymbol{\mu}}{2}\|\boldsymbol{G}\boldsymbol{x} - \boldsymbol{u}\|_2^2, \quad (4)$$

where $\boldsymbol{\mu}$ is the Lagrange multiplier. The above minimization problem can be now fitted into the ADMM framework and subsequently decoupled into the following two subproblems:

Subproblem 1: $\boldsymbol{x}_{k+1} = \arg\min_{\boldsymbol{x}} f_1(\boldsymbol{x}) + \frac{\boldsymbol{\mu}}{2}\|\boldsymbol{G}\boldsymbol{x} - \boldsymbol{u}_k - \boldsymbol{d}_k\|_2^2$
Subproblem 2: $\boldsymbol{u}_{k+1} = \arg\min_{\boldsymbol{u}} f_2(\boldsymbol{u}) + \frac{\boldsymbol{\mu}}{2}\|\boldsymbol{G}\boldsymbol{x}_{k+1} - \boldsymbol{u} - \boldsymbol{d}_k\|_2^2$
Updating: $\boldsymbol{d}_{k+1} \leftarrow \boldsymbol{d}_k - (\boldsymbol{G}\boldsymbol{x}_{k+1} - \boldsymbol{u}_{k+1})$.

All that remains are to solving the two subproblems, for which we demonstrate that they can be elegantly solved using standard methods after simple algebraic transformation in Additional file 1: S.2. For the clarity of the numerical scheme, a short introduction of ADMM is also provided in Additional file 1: S.1.

Parameter pruning

The dispersion parameter α is associated with the negative binomial distributions of the different CN states. In our experiments, the dispersion parameter was estimated by MLE. In simulation experiments, the RDs from simulated reads of the chromosome-21 sequence without implanted CNVs were used for the MLE estimation of α. In empirical experiments, the RD signals from a normal single cell under accession number SRR052047 were employed for estimation of α. The parameter λ_1 is used to penalize the total variational term, and λ_2 is used to control the sparsity of the recovered signal. Both of the two parameters were estimated by trials on preliminary experiments. The copy number duplications were implanted artificially in the RD data of SRR052047 by adding one CN to any bins with the duplications. The copy number deletions were generated similarly by subtracting one to any bins overlapped the deletions. It should be noted that SRR052047 was considered as a clean sample (CN=2) and thus its copy number status was known. We run the simluation experiments for different values of λ_1 and λ_2. The Euclidean distance between the fitter signals and the real copy number signals was calculated for evaluating the CNV detection performance. As shown in Fig. 3, one may observe that when the λ_1 was set as 1 and λ_2 was set as 1 the Euclidean distance achieved the minimum. For real data experiments, the two parameters of λ_1 and λ_2 were pruned around 1.

Results and discussion

Simulation experiments

To evaluate the performance of nbCNV, experiments on a simulation dataset from a chromosome sequence with

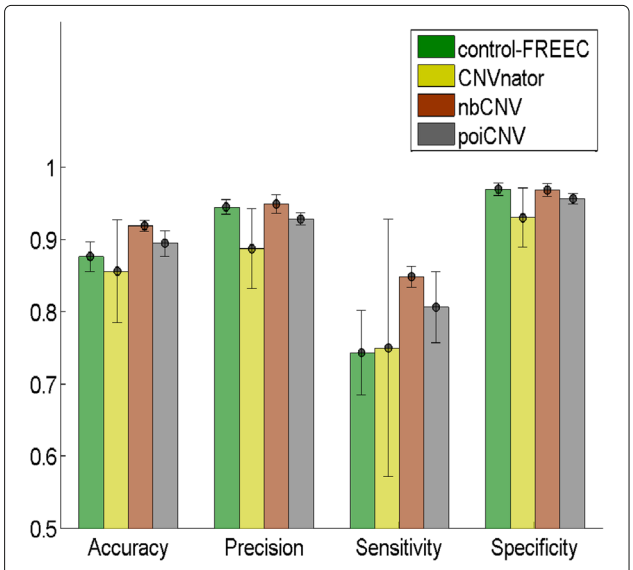


Fig. 5 The performance of implanted CNVs detection after four methods of control-FREEC, CNVnator, nbCNV and poiCNV at chromosome-21. Each bar in the plot represents the mean based on 100 simulations of the corresponding measurement as determined by each method. In addition to the error bars, the nbCNV method can clearly be seen to have achieved superior performance in terms of accuracy, precision and sensitivity

implanted CNVs were conducted. The chromosome 21 of GRCh37 was used as a template. Variants including duplications and deletions were randomly implanted into it. Our experiments considered only duplications (CN = 3) and deletions (CN = 1) since these two types of CNVs are typically the most challenging problem in distinguishing them from normal CNs. We first doubled the chromosome-21 sequence (CN = 1) to generate the diploid sequence (CN = 2). The length of chromosome 21 without unknown sequences is 35106692 bp and the size of CNVs ranged from 300000 bp to 2000000 bp. For each simulation, 10 CNVs were artificially implanted into the chromosome, from which simulated single-end sequencing reads were created by WgSim [18]. WgSim is a simulation tool to create NGS reads, including single nucleotide polymorphisms, insertion-deletions and sequencing errors from a reference sequence. The simulated reads by WgSim were further contaminated by noises following negative binomial distribution to mimic the technical noises introduced by amplification and sequencing [12]. A total number of 160452 reads with coverage of 0.22× were generated. Each single-end reads is in 50-bp, similar to the Illumina sequencing platform. As shown in Fig. 4, the frequency histogram of the simulated RCs was approximated by negative binomial and Poisson distribution. One may note that the frequency of the simulated reads was nicely characterized by the negative binomial distribution.

To have quantitative comparison on CNV detection, the sampled short reads were aligned back to the reference sequence by bowtie2 [16]. Its output sam files were used as the input for control-FREEC [6], CNVnator [1], nbCNV and our earlier work based on Poisson model named by poiCNV) [32] for performance comparison. As for control-FREEC, the chromosome-21 sequence with implanted CNVs were served as the control sequence. The bin size used in control-FREEC and CNVnator analyses was 50000. The above simulation was run 100 times independently. For quantitative comparison, four measurements including accuracy, precision, sensitivity and

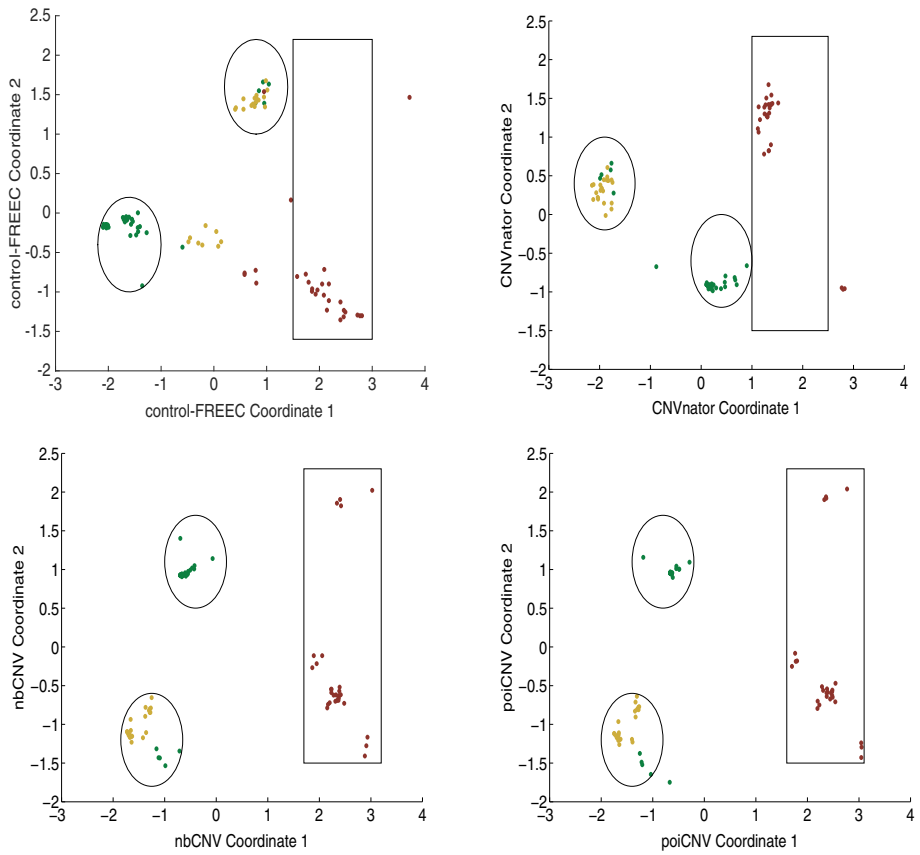


Fig. 6 Multidimensional scaling of 100 single cells. Diploid (2N), hypodiploid (1.7N) and aneuploid (3N or 3.3N) fractions are shown in green, yellow and red, respectively. Clustering results after nbCNV were better than those after CNVnator, when comparing the number of covered dots. Compared with the other three methods, control-FREEC resulted in a smaller inter-cluster distance between diploid and hypodiploid fractions, and thus was less satisfactory

specificity were recorded and calculated. Their definitions are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
$$Precision = \frac{TP}{TP + FP}$$
$$Sensitivity = \frac{TP}{TP + FN}$$
$$Specificity = \frac{TN}{TN + FP}$$

where true positive (TP) is the total number of instances when the CNV regions are correctly identified and true negative (TN) is the number of instances when the normal regions (CN = 2) are detected properly. False positive (FP) and false negative (FN) are defined similarly. The experimental results are summarized in Table 1, in which the best value was highlighted in bold. For visual comparison, a bar graph was also drawn in Fig. 5. Among the 100 simulations, the nbCNV performed superior to poiCNV and its peers by achieving the highest measurements of accuracy, precision and sensitivity. Moreover, nbCNV also resulted in a smaller standard deviations than by control-FREEC and CNVnaotr. Compared with control-FREEC and CNVnaotr, the superior performance of nbCNV is attributed to its effective data preprocessing and robustness in parameter pruning. Compared with poiCNV, the nice performance of nbCNV is attributed to its appropriate noise modelling.

Application to SCS data from 100 single cells

To further assess the performance of nbCNV in real applications, a SCS dataset from 100 single cells was downloaded from the NCBI SRA under accession number SRP002535 and tested. The original samples were selected from high-grade (III) triple-negative (ER^-, PR^-, $HER2^-$) ductal carcinomas (T10) [21]. They were preprocessed by flow sorting of single nuclei, whole genome amplification, library construction, and finally sequenced on an Illumina Genome Analyzer [3]. The 100 Illumina runs generated a total of 1.1×10^9 reads, 5.8×10^{10} base calls (33.3 Gb downloads in sra format) and were thus of low coverage. The data has been used to study the evolutionary dynamics and population structure of tumors in order to have a comprehensive view of the evolutionary process occurring in individual tumor cells [21]. They have been analyzed by fluorescence-activated cell sorting and therefore their ploidy levels were known, including 47 diploids or pseudodiploids (2N), 24 hypodiploids (1.7N) and 29 aneuploids (3N or 3.3N). The diploids or

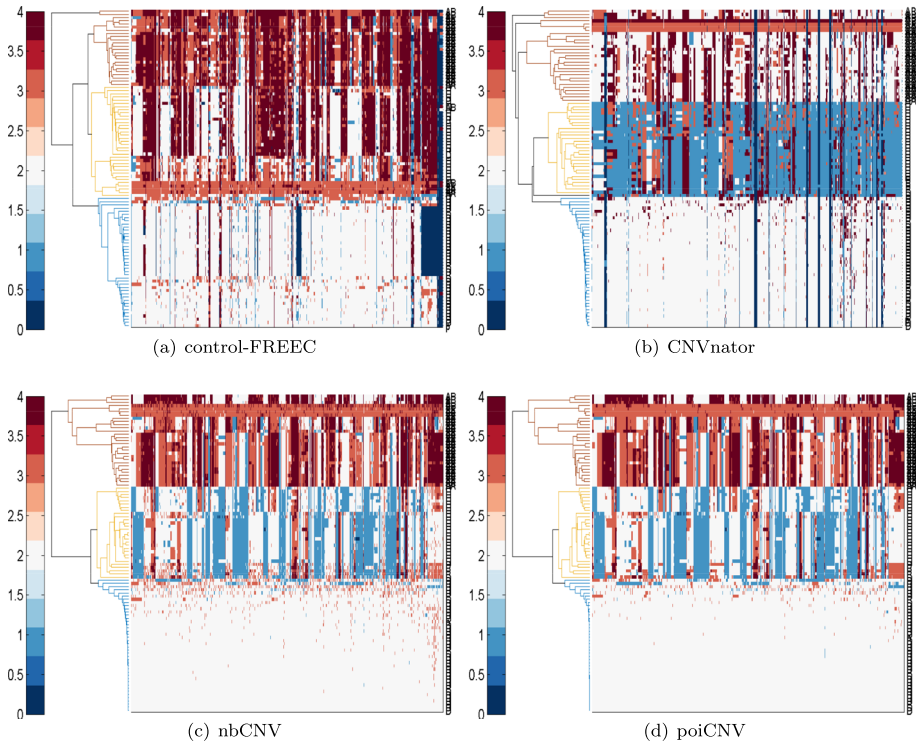


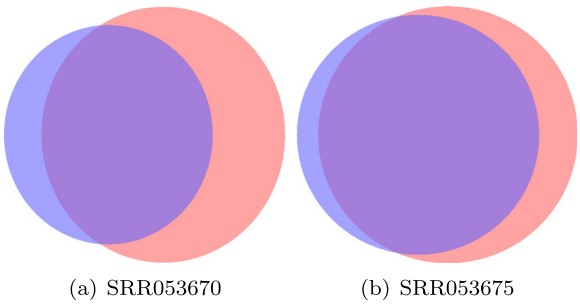
Fig. 7 Results of hierarchical clustering and heat map of copy number profiles of 100 single cells after **a** control-FREEC, **b** CNVnator, **c** nbCNV and **d** poiCNV analysis. Different colors are used to distinguish the generated clusters. A comparative evaluation of the performance of the different detection methods based on the clustering results is given in Table 2

pseudodiploids part consists of cells which have a small number of CNVs as a whole, while the hypodiploids part shows narrow deletions and the aneuploids part shows numerous copy number duplications [21]. The ploidy information could serve as benchmark information for evaluating the clustering performance in these 100 cells.

The proposed nbCNV as well as the other three methods of poiCNV, CNVnator and control-FREEC were applied on the sequence data to have their CN profiles. For visual comparison, multidimensional scaling (MDS) was performed by mapping each sample from the high-dimensional space to a visually acceptable one (i.e., two dimensions). The diploid (2N), hypodiploid (1.7N) and aneuploid (3N or 3.3N) fractions are highlighted in Fig. 6. Five diploid cells were mistakenly classified into the hypodiploid fraction by nbCNV and poiCNV. However, when we retrospectively examined the five diploid cells, we found that they possess abundant CNVs compared with other diploid cells and thus making it difficult to be merged with others. The clustering result on the CN profiles after nbCNV was better than the one by CNVnator by covering most of the sample dots. The one after control-FREEC method resulted in a smaller inter-cluster distance between diploid and hypodiploid fractions. To further visualize the evolutionary history of the 100 single cells, hierarchical clustering was computed and shown in Fig. 7. The misclassified cell number after each method was summarized in Table 2. The proposed nbCNV and poiCNV achieved superior performance by only misclassifying five cells. In comparison, CNVnator misclassified 6 cells and control-FREEC misclassified 10 cells. It should be noted that although ploidy information could serve as benchmark information for evaluating the clustering results. However, using clustering accuracy to evaluate the performance of methods was coarse-grained. For example, nbCNV and poiCNV were shown to perform equally well in term of MDS and clustering accuracy. Since the real copy number profiles of these sequence data were unknown, we pictured the number of detected CNVs after nbCNV and poiCNV by Wayne chart at two typical cells (SRR053670, SRR053675). As shown in Fig. 8, nbCNV can detect more CNVs than by poiCNV. It implies a possible good coverage yet higher false positive. We also reported the detection results on other cells by nbCNV and poiCNV in Additional file 1: S.3.

Table 2 Quantitative evaluation of the four copy number variant detection methods based on clustering of single-cell sequencing data from 100 cells

Methods	Classification error count		
	cluster 1	cluster 2	cluster 3
CNVnator	0	6	0
nbCNV	0	5	0
control-FREEC	0	10	0
poiCNV	0	5	0



(a) SRR053670 (b) SRR053675

Fig. 8 Wayne comparison on detected CNVs by nbCNV and poiCNV on two samples. The pink part measures the number of CNVs detected by nbCNV while the purple represents poiCNV and the overlay region represents both. It implies of a possibly better coverage yet higher false positive by nbCNV

Conclusions

We have presented a RD-based method to detect CNVs from over-dispersed sequencing data such as SCS data. Taking into account the over-dispersed noise in the SCS data and the characteristics of CNV patterns, the method uses negative binomial distributions to model the RD signal and imposes sparsity and smoothness constraints to transform CNV detection into a quadratic optimization problem. Comparative experiments with other CNV detection methods on simulated data and an empirical SCS dataset demonstrated that our method is superior in terms of accuracy, precision and sensitivity for CNV detection. Compared with other methods, the robustness in parameter pruning in our CNV detection method contributes to a more steady performance.

Additional file

Additional file 1: Contains a formal description of the ADMM algorithm and the CNV detection results on SCS ductal carcinomas (T10) data by nbCNV and poiCNV. (PDF 166 kb)

Abbreviations

CNVs: Copy number variants; CN: Copy number; SCS: Single-cell sequencing; WGA: Whole-genome amplification; RD: Read depth; GC: Guanine-cytosine; CBS: Circular binary segmentation; ADMM: Alternating direction minimization method; MAPD: Multiple absolute pairwise difference

Funding

This work was partially supported by the National Natural Science Foundation of China (61372141), Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase), Science and Technology Planning Project of Guangdong Province, Open Funds of State Key Laboratory of Oncology in South China, and the Fundamental Research Fund for the Central Universities (2015ZZ025).

Availability of data and materials
The datasets supporting the conclusions of this article are available in the NCBI Sequence Read Archive (SRA) repository [accession number SRP002535 and http://www.ncbi.nlm.nih.gov/sra/SRX021401[accn]]. The relevant code can be downloaded from https://github.com/zcszcstl/nbCNV.

Authors' contributions
HMC provided method, design, development and authorship. CSZ designed the experiment and evaluated the performance and drafted the manuscript. JYH provided advisement on experiments. YS provided the data preparation and manuscript editing. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
The authors declare that ethics approval and consent to participate are not applicable to this study.

Received: 3 March 2016 Accepted: 4 September 2016
Published online: 17 September 2016

References

1. Abyzov A, Urban AE, Snyder M, Gerstein M. Cnvnator: an approach to discover, genotype, and characterize typical and atypical cnvs from family and population genome sequencing. Genome Res. 2011;21(6):974–84.
2. Amarasinghe KC, Li J, Halgamuge SK. Convex: copy number variation estimation in exome sequencing data using hmm. BMC Bioinformatics. 2013;14(Suppl 2):S2.
3. Baslan T, Kendall J, Rodgers L, Cox H, Riggs M, Stepansky A, Troge J, Ravi K, Esposito D, Lakshmi B, Wigler M, Navin N, Hicks J. Genome-wide copy number analysis of single cells. Nat Protoc. 2012;7(6):1024–41.
4. Baslan T, Kendall J, Ward B, Cox H, Leotta A, Rodgers L, Riggs M, D'Italia S, Sun G, Yong M, et al. Optimizing sparse sequencing of single cells for highly multiplex copy number profiling. Genome Res. 2015;25(5):714–24.
5. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, Sboner A, Esgueva R, Pflueger D, Sougnez C, et al. The genomic complexity of primary human prostate cancer. Nature. 2011;470(7333): 214–20.
6. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, Janoueix-Lerosey I, Delattre O, Barillot E. Control-freec: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics. 2012;28(3):423–5.
7. Cai X, Evrony GD, Lehmann HS, Elhosary PC, Mehta BK, Poduri A, Walsh CA. Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. Cell Rep. 2014;8(5):1280–9.
8. Carén H, Kryh H, Nethander M, Sjöberg R-M, Träger C, Nilsson S, Abrahamsson J, Kogner P, Martinsson T. High-risk neuroblastoma tumors with 11q-deletion display a poor prognostic, chromosome instability phenotype with later onset. Proc Natl Acad Sci. 2010;107(9):4323–8.
9. Chiang DY, Getz G, Jaffe DB, O'Kelly MJ, Zhao X, Carter SL, Russ C, Nusbaum C, Meyerson M, Lander ES. High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat Methods. 2009;6(1):99–103.
10. 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 Bioinformatics. 2013;14(1):150.
11. Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, Zhang H, Estes A, Brune CW, Bradfield JP, et al. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature. 2009;459(7246):569–73.
12. Grün D, Kester L, van Oudenaarden A. Validation of noise models for single-cell transcriptomics. Nat Methods. 2014;11(6):637–40.
13. Handyside AH, Robinson MD, Simpson RJ, Omar MB, Shaw M-A, Grudzinskas JG, Rutherford A. Isothermal whole genome amplification from single and small numbers of cells: a new era for preimplantation genetic diagnosis of inherited disease. Mol Hum Reproduction. 2004;10(10):767–72.
14. Klambauer G, Schwarzbauer K, Mayr A, Clevert D, Mitterecker A, Bodenhofer U, Hochreiter S. 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;40(9):e69.
15. Krepischi A, Achatz M, Santos E, Costa SS, Lisboa B, Brentani H, Santos TM, Gonċalves A, Nóbrega AF, Pearson PL, et al. Germline dna copy number variation in familial and early-onset breast cancer. Breast Cancer Res. 2012;14(1):R24.
16. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9(4):357–59.
17. Lasken RS. Single-cell genomic sequencing using multiple displacement amplification. Curr Opin Microbiol. 2007;10(5):510–6.
18. Li H. wgsim-read simulator for next generation sequencing. 2013. https://github.com/lh3/wgsim.
19. Magi A, Tattini L, Pippucci T, Torricelli F, Benelli M. Read count approach for dna copy number variants detection. Bioinformatics. 2012;28(4):470–8.
20. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nat Methods. 2009;6(11):S13–20.
21. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, Muthuswanmy L, Kransnitz A, McCombie W, Hicks J, Wigler M. Tumour evolution inferred by single-cell sequencing. Nature. 2011;472(7341):90–4.
22. Navin NE. Cancer genomics: one cell at a time. Genome Biol. 2014;15:452.
23. Ng MK, Weiss P, Yuan X. Solving constrained total-variation image restoration and reconstruction problems via alternating direction methods. SIAM J Sci Comput. 2010;32(5):2710–36.
24. Nilsen G, Liestøl K, Van Loo P, Vollan HKM, Eide MB, Rueda OM, Chin S-F, Russell R, Baumbusch LO, Caldas C, et al. Copynumber: Efficient algorithms for single-and multi-track copy number segmentation. BMC Genomics. 2012;13(1):591.
25. Olshen AB, Venkatraman E, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based dna copy number data. Biostatistics. 2004;5(4):557–72.
26. Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. Nat Genet. 2005;37:S11–S17.
27. Seshan VE, Olshen A. Dnacopy: Dna copy number data analysis. 2011. http://www.bioconductor.org/packages/.
28. Silander K, Saarela J. Whole genome amplification with phi29 dna polymerase to enable genetic or genomic analysis of samples of low dna yield. In: Genomics Protocols. Springer; 2008. p. 1–18.
29. Steinberg S, de Jong S, Mattheisen M, Costas J, Demontis D, Jamain S, Pietiläinen OP, Lin K, Papiol S, Huttenlocher J, et al. Common variant at 16p11. 2 conferring risk of psychosis. Mol Psychiatry. 2014;19(1):108–14.
30. Wells D, Sherlock JK, Delhanty JD, Handyside AH. Detailed chromosomal and molecular genetic analysis of single cells by whole genome amplification and comparative genomic hybridisation. Nucleic Acids Res. 1999;27(4):1214–8.
31. Xie C, Tammi MT. Cnv-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinformatics. 2009;10(1):80.
32. Xu B, Cai H, Zhang C, Yang X, Han G. Copy number variants calling for single cell sequencing data by multi-constrained optimization. Comput Biol Chem. 2016.
33. Zhang Z, Lange K, Ophoff R, Sabatti C. Reconstructing dna copy number by penalized estimation and imputation. Ann Appl Stat. 2010;4(4):1749.
34. Zhang Z, Lange K, Sabatti C. Reconstructing dna copy number by joint segmentation of multiple sequences. BMC Bioinformatics. 2012;13(1):205.
35. Zhou X, Yang C, Wan X, Zhao H, Yu W. Multisample acgh data analysis via total variation and spectral regularization. IEEE/ACM Trans Computat Biol Bioinform. 2013;10(1):230–5.
36. Zong C, Lu S, Chapman AR, Xie XS. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science. 2012;338(6114):1622–6.