Bioinformatics, 2019, 1–8
doi: 10.1093/bioinformatics/bty1041
Advance Access Publication Date: 14 January 2019
Original Paper

Genome analysis

An accurate and powerful method for copy number variation detection

Feifei Xiao[1,*], Xizhi Luo[1], Ning Hao[2], Yue S. Niu[2], Xiangjun Xiao[3], Guoshuai Cai[4], Christopher I. Amos[3] and Heping Zhang[5,*]

[1]Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, SC 29201, USA, [2]Department of Mathematics, University of Arizona, Tucson, AZ 85721, USA, [3]Department of Quantitative Sciences, Baylor College of Medicine, Houston, TX 77030, USA, [4]Department of Environmental Health Science, University of South Carolina, Columbia, SC 29201, USA and [5]Department of Biostatistics, Yale School of Public Health, New Haven, CT 06510, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on September 6, 2018; revised on November 28, 2018; editorial decision on December 18, 2018; accepted on January 9, 2019

Abstract

Motivation: Integration of multiple genetic sources for copy number variation detection (CNV) is a powerful approach to improve the identification of variants associated with complex traits. Although it has been shown that the widely used change point based methods can increase statistical power to identify variants, it remains challenging to effectively detect CNVs with weak signals due to the noisy nature of genotyping intensity data. We previously developed modSaRa, a normal mean-based model on a screening and ranking algorithm for copy number variation identification which presented desirable sensitivity with high computational efficiency. To boost statistical power for the identification of variants, here we present a novel improvement that integrates the relative allelic intensity with external information from empirical statistics with modeling, which we called modSaRa2.

Results: Simulation studies illustrated that modSaRa2 markedly improved both sensitivity and specificity over existing methods for analyzing array-based data. The improvement in weak CNV signal detection is the most substantial, while it also simultaneously improves stability when CNV size varies. The application of the new method to a whole genome melanoma dataset identified novel candidate melanoma risk associated deletions on chromosome bands 1p22.2 and duplications on 6p22, 6q25 and 19p13 regions, which may facilitate the understanding of the possible roles of germline copy number variants in the etiology of melanoma.

Availability and implementation: http://c2s2.yale.edu/software/modSaRa2 or https://github.com/FeifeiXiaoUSC/modSaRa2.

Contact: xiaof@mailbox.sc.edu or heping.zhang@yale.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Copy number variations (CNVs) are deletions (<2) or duplications (>2) in DNA copies at a specific chromosomal location in the genome. According to their origin, CNVs are usually classified into two categories. Germline CNVs refer to inherited variants, often existing as polymorphisms at the population level and may explain part of the 'missing heritability' (Maher, 2008). Copy number aberrations in somatic cells, referred to as somatic CNVs, have also been investigated to understand the non-inherited component of diseases, especially in tumorigenesis (Qiu *et al.*, 2017). Up until now, studies have provided evidence to support the unique roles of CNVs in the etiology of many diseases such as cancer (Chen *et al.*, 2013a,b;

Kumaran *et al.*, 2017; Lin *et al.*, 2011), autoimmune (Li *et al.*, 2017; Marshall *et al.*, 2017) and neurological diseases (Hou *et al.*, 2013; Stuart *et al.*, 2012). For example, copy number gain of beta-defensin genes has been revealed to be associated with increased risk of psoriasis in three cohorts of European origin (Barnes *et al.*, 2008; Hollox *et al.*, 2008; Stuart *et al.*, 2012). Also, the deletions of complement 4a and 4b have been found to increase risk for autoimmune diseases (Li *et al.*, 2017).

Although numerous statistical approaches have been developed for different platforms (e.g. SNP array, exome sequencing), it remains persistently difficult to accurately identify CNVs. One complication is the irregularity of CNV occurrence since they do not occur in the same location across individuals. Another complication is the existence of random noise in the data, which easily leads to invalid CNV calling. Also, different genotyping platforms do not share methods due to their technological complications. Well-developed array-based CNV analytical tools are usually based on segmentation and smoothing of Log R Ratio (LRR) and B-allele frequency (BAF), which integrate evidence for copy number status (Wang *et al.*, 2007). LRR measures the normalized total intensity of the possible alleles for a given marker, from which the magnitude of mean change, referred to as jump size, are used for inference of breakpoints. The larger the jump size, the more likely the existence of CNV. BAF is the normalized measure of relative signal intensity ratio of one of the possible alleles, a variation of which from expected signals for a diploid region reflects the underlying copy number states. Some existing statistical tools for sequencing generated data have already used the information embedded in BAF in the segmentation. For example, SomatiCA was proposed to quantify somatic copy number aberrations by integrating the read counts sequencing data with the information from BAF (Chen *et al.*, 2013a,b). PennCNV, based on a Hidden Markov model, is another widely used method integrating both intensities (Wang *et al.*, 2007). However, with change-point segmentation methods, the relative allelic intensity information is still underutilized by many statistical models, and often is integrated in an *ad hoc* manner.

For change point methods, the main goal is finding multiple breakpoints in the expansive chromosome, where the length is typically thousands of SNPs. A previous review (Zhang, 2010) provided a thorough introduction to the application of change-point models in CNV detection. Among them, circular binary segmentation (CBS) is a change-point test applied recursively to determine all of the breakpoints (Olshen *et al.*, 2004). As a default segmentation algorithm, CBS has been widely implemented into CNV detection software and tools such as CNV workshop, SegGene and NEXUS (Darvishi, 2010; Deng, 2011; Gai *et al.*, 2010). It provides a very consistent performance although it presents high computational complexity. A more recent result showed the improved computational speed; however, the heavy computation still presents obstacles for its wide application in high dimensional data with large sample sizes (Venkatraman and Olshen, 2007). Also, issues arise from the application of these methods to genetic studies such as inflation of false positives especially when the jump size is small.

To address the issue of computational complexity and other application challenges, we previously developed a robust and powerful method based on a normal mean change-point model, modSaRa, implemented for analyzing array-based data (Xiao *et al.*, 2015). With modSaRa, we adopted a local search strategy in the segmentation which largely increased the computational efficiency. It provides a user-friendly tool that identifies CNVs across the genome with optimal sensitivity and specificity. However, inflation of false positives is still a challenge, especially when the CNV signal is weak.

In this study, we therefore fully utilized the relative allelic intensity, BAF, in the segmentation algorithm to boost the power, along with integrating external empirical statistics information to efficiently control the false discovery rate. We then implemented this improved segmentation algorithm and applied it to the modSaRa2 software, for the analysis of array-based data. Our simulation studies illustrated the increased power and well-controlled false discovery rate of modSaRa2, especially for detecting CNVs with weak signals or large CNVs. The computational speed is very fast, which is about 9 s for processing a chromosome with 90 000 markers. We also applied this new method to a whole genome cutaneous melanoma study to uncover the roles of germline CNVs in the etiology of melanoma.

2 Materials and methods

2.1 Overview of modSaRa

First, we will introduce the background of modSaRa, which was proposed for analyzing microarray data. We start from the screening and ranking algorithm (SaRa) from the viewpoint of hypothesis testing and in the context of CNV detection. Assume that we have a sequence or more specifically, a chromosome, let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ be the random variables of genetic intensities with n markers (e.g. SNP marker). For the ith marker, we assume

$$y_i = \mu_i + \varepsilon_i, \;\; i = 1, \;\; 2, \ldots, n. \tag{1}$$

$\boldsymbol{\mu} = (u_1, \ldots, u_n)^T$ is the underlying mean; the errors ε_i are *i.i.d* and follow $N(0, \sigma^2)$. A change point is a position τ such that $\mu_\tau \neq \mu_{\tau+1}$. In this model, our goal was to simultaneously make inference on the existence and location of τ's. Therefore, it can be stated as the hypothesis testing problem of multiple change points:

$$H_0 : \mu_1 = \cdots = \mu_n, \;\; \text{versus},$$

$$H_1 : \mu_1 = \cdots = \mu_j \neq \mu_{j+1} = \cdots = \mu_k \neq \mu_{k+1} = \cdots = \mu_n \text{ for some } j \text{ and } k.$$

To test the alternative hypothesis of multiple change points, a locally defined diagnostic statistic was therefore proposed for screening the whole sequence as

$$D_h(j) = \left(\sum_{l=1}^{h} y_{j+1-l} - \sum_{l=1}^{h} y_{j+l} \right) \Big/ h, \quad j = h+1, h+2, \ldots, n-h, \tag{2}$$

where h is a fixed integer representing bandwidth. The above function is simply evaluating the difference between the points on left side and those on right side of point j, with window size $2h$. Obviously, the magnitude of $D_h(j)$ reflects if the point is near or is a change-point. The advantage of this strategy is its usages of local information while considering the global sequence. We can consider a more general form of local diagnostic statistic, which is the weighted average of y_i's near the point of interest j,

$$D_h(j) = \sum_{i=1}^{n} w_i(j) y_i, \tag{3}$$

where the weight function $w_i(j) = 0$ when $|i - j| > h$. In Equation (2), we have an equal-weight diagnostic function

$$w_i(j) = \begin{cases} \frac{1}{h} & 1 - h \leq i - j \leq 0 \\ -\frac{1}{h} & 1 \leq i - j \leq h \\ 0 & \text{otherwise} \end{cases}. \tag{4}$$

Notice that in the above function, each point had an equal chance of being a change point.

To address issues arising from the application of SaRa, we further proposed a robust and powerful method, modSaRa, as an endeavor to detect CNVs with SNP array data (Xiao *et al.*, 2015). Specifically, we proposed using multiple bandwidths in the screening and ranking steps to optimize the sensitivity in segmentation, then a Gaussian mixture model-based clustering strategy to optimize its specificity by simultaneously removing false positives and clustering CNV segments. The Expectation-Maximization algorithm (Dempster *et al.*, 1977) was used in the clustering step by assuming the segmentation means were following normal distributions. By simulations, we have illustrated the greater accuracy and computational speed of the modSaRa over CBS with SNP array data. Details can be found in the previous method and package paper (Xiao *et al.*, 2015, 2017). For algorithms involving bandwidths, the value of bandwidths should be carefully chosen (Niu and Zhang, 2012). In modSaRa2, we suggest using three different bandwidths with the choices depending on the applications (for example, 5, 10, 15 for microarray data).

2.2 Gaussian likelihood copy number estimation

Although we made important progress with modSaRa, our previous work still had limitations. First, the segmentation algorithm implemented in modSaRa was over-sensitive, resulting in abundant false positives in the calling results. Second, the performance of the test can be optimized by taking prior information of the genetic intensities such as empirical statistics from external sources. To address the above issues, we here have developed a Gaussian likelihood copy number estimate to efficiently integrate the prior empirical statistics into the statistical modeling, which efficiently improves the overall accuracy.

We consider a diallelic locus A on a chromosome with two alleles, A_1 and A_2. Consequently, given a normal diploid state when no duplication or deletion occurs, the genotypes include A_1A_1,A_1A_2 and A_2A_2. When an allele is gained, the genotype can be $A_1A_1A_1$, $A_1A_1A_2$, $A_1A_2A_2$, $A_2A_2A_2$ depending on the original genotype and which allele is duplicated. Naturally, the genotypes will be A_1 or A_2 when one copy of the alleles is lost. The raw signal intensity values measured for the A_1 and A_2 alleles are then subject to a five-step normalization procedure using the signal intensity of all SNPs. The procedure produces the normalized intensity values, X_1 and X_2, for the two alleles respectively. As a normalized measure of total signal intensity, the log R Ratio (LRR) value for each SNP is calculated as $\mathrm{LRR} = \log_2(R_{\mathrm{observed}}/R_{\mathrm{expected}})$, where $R = X_1 + X_2$ refers to the total signal intensity and R_{expected} is computed from linear interpolation of canonical genotype clusters (Peiffer *et al.*, 2006). BAF is another dimension of measure representing normalized measure of relative signal intensity of the minor allele (e.g. A_2). The detailed mathematical definition can be found in previous published literature (Wang *et al.*, 2007). Briefly, when the segments are normal state (diploids), the BAF values usually display three clusters at values of 0, 0.5 and 1 with small variance, representing homozygotes of A_1 allele (A_1A_1), heterozygotes (A_1A_2) and homozygotes of A_2 allele (A_2A_2), respectively. When deletion occurs, only one allele is left, hence the BAF means will be 0 and 1, respectively.

To integrate the information from BAF, we first introduce a less sparse intensity compared to BAF, Lesser Allele Frequency (LAF):

$$LAF = \begin{cases} BAF & \forall BAF \leq 0.5 \\ 1 - BAF & \forall BAF > 0.5 \end{cases}. \tag{5}$$

The above transformation makes the modeling more efficient since we usually do not need to distinguish certain genotype pairs such as $A_1A_1A_2$ or $A_1A_2A_2$ when on copy is gained. Given a diallelic locus, let $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^T$ represent the sequence of LAF; $\boldsymbol{G}^* = (G_1^*, \ldots, G_n^*)^T$ denote the genotype of a locus with eight possible genotypes; $\boldsymbol{S}^* = (S_1^*, \ldots, S_n^*)^T$ be the copy number state that needs to be estimated, which includes deletion of double copy (Del.D), deletion of single copy (Del.S), normal state (Diploids), duplication of single copy (Dup.S) and duplication of double copy (Dup.D). The connection between genotypes and these copy number states can be found in Supplementary Table S1. Among these states, two-copy duplications/deletions (Del.D and Dup.D) have relatively strong CNV signals and therefore can be easily captured, whereas the single-copy changes (Del.S and Dup.S) are more difficult to identify.

For the ith locus, we assume both LRR and LAF follow normal distributions conditional on the genotype, that $Y_i \sim N(\mu_Y, \sigma_Y^2|G_i^*)$ and $Z_i \sim N(\mu_Z, \sigma_Z^2|G_i^*)$. The mean μ and variance σ^2 are unknown parameters and will be estimated from the external information of empirical statistics for LRR and LAF, respectively. For the estimation of these parameters, cnvPartition embedded in the Genome Studio software developed by Illumina (https://www.illumina.com/documents/products/technotes/technote_cnv_algorithms.pdf), provides a good summary of all the genotypes and copy number states with the estimated mean and variance for each state (Supplementary Table S2). For data generated from the Affymetrix platform, the empirical means are provided in the Affymetrix website, whereas the variance can be easily estimated from the samples. Owing to the normalization procedures of the intensities, when both copies are deleted, the BAF is generated as a random value. As a result, $Z_i \sim Uniform(0,\ 0.5)$ for the copy number state of Del.D. For the locus i, the likelihood of being each hidden genotype is given by:

$$\begin{aligned} L_{G_i} = f_N(Y_i|\mu_Y, \sigma_Y, G_i^*) f_N(Z_i|\mu_Z, \sigma_Z, G_i^*)^{1-I(G_i^*=DD)} \\ f_U(Z_i|\mu_Z, \sigma_Z, G_i^*)^{I(G_i^*=DD)}, \end{aligned} \tag{6}$$

$I(G_i^* = DD)$ is the indicator function which equals to one when the genotype is Del.D. These genotype likelihoods are then summarized by five composite copy number likelihoods by $L_{s_k,i} = \sum_{G_i \subset s_k} L_{G_i} (k = 1,\ 2,\ \ldots,\ 5)$ with respect to the kth copy number state, where $s = \{Del.D,\ Del.D, Diploid, Dup.S, Dup.D\}$ representing the assumed five copy number states.

The preliminary copy number estimate eCN_i is therefore defined as

$$eCN_i = \frac{\sum_{k=1}^{5} (k-1) L_{s_k,i}}{\sum L_{s_k,i}}. \tag{7}$$

The expected value of the continuous variable ***eCN*** will be integer values from 0 to 4 reflecting the five copy number states, respectively.

2.3 CNV detection and post CNV-calling steps

After we obtained the preliminary copy number estimates ***eCN***, the screening and ranking algorithm with multiple bandwidths was applied for breakpoints identification. The goal was to identify regions of the genome where the values of ***eCN*** are consistently higher or lower than two, the expected value of a diploid segment. Then the procedure of change point identification and copy number assignments was conducted in a similar manner as in modSaRa (Supplementary Methods and Results A1). A false discovery rate approach was used to adjust the P-values for the multiple comparison

problem (Supplementary Methods A2). For the subset selection of candidate change points, the threshold of *P*-values is an important user-defined parameter in the package. Using high quality CNV calls from three previous studies as validation sets (Conrad *et al.*, 2010; International HapMap *et al.*, 2010; McCarroll *et al.*, 2008), we evaluated the performance of the modSaRa2 package under different *P*-value thresholds. As a result, we recommended the significance level to be 0.01 or 0.05 in applications (Supplementary Methods and Results A3). By default, the significance level was set at a more stringent level at 0.01 in the package. Some post CNV calling quality control steps were also developed for reliable CNV calling, including an *ad hoc* procedure for removing false positives and merging adjacent CNV calls (Supplementary Methods and Results A4). Supplementary Figure S1 systematically illustrates how modSaRa2 works.

2.4 Software implementation

In the current modSaRa2, instead of calculating the sequence of empirical values of reverse cumulative distribution value of $D_b(j)$ from a sequencing with no change-points recurrently as in modSaRa, we calculated once beforehand and repeatedly used it. Consequently, we reduced computational burden. For example, the computational time was dramatically saved, which is 8.63 s for a chromosome with 86 694 markers. The above calculation is based on analysis made at a desktop workstation with an Intel Core i5 CPU 2.6 GHz processor and 8.0 GB of RAM. The modSaRa package is now publicly available at: http://c2s2.yale.edu/software/modSaRa2 and https://github.com/FeifeiXiaoUSC/modSaRa2.

2.5 Numerical simulations

We first simulated data to evaluate the performance of modSaRa2 as compared to the previous version of modSaRa and other state-of-the-art methods. We used the simulations mimicking the data generated from Illumina platform. In total, fifty samples or sequences with 10 000 markers per sequence were generated assuming independence of the markers. In each sequence, twenty dispersed and non-overlapping CNVs were generated, the locations of which were randomly selected but shared by individuals. To minimixze detection bias, the distance gap between two adjacent CNVs was more than 10 markers. For each simulation setting, there were 1000 CNVs or equivalently 2000 change points in total.

We characterized the performance of modSaRa2 in different scenarios of CNV lengths and states. The CNV lengths vary among 10–50, 50–100 or 100–200 markers. The copy number states included Dup.S, Dup.D, Del.S and Del.D. For each scenario, the LRR and BAF intensities were generated from a normal distribution with the mean (μ) and standard deviation (σ) as those provided with the values computed by Illumina (Supplementary Table S2). Let us take the scenario of Del.S and CNV length of 10–50 as an example. First, the length of copy number segments for the independent CNVs was randomly generated from values between 10 and 50. Then the genomic markers within each CNV region received values generated from distribution with $N(-0.45, 0, 18^2)$ for LRR intensities. For Del.S, the genotype can be either A_1 or A_2, which resulted in randomly generated values of 0 or 1 with variance setting at 0.03 in BAF values (Supplementary Table S1). For the rest of the same sequence with diploids (normal state), the LRRs were generated from $N(0, 0, 18^2)$ whereas the BAFs were generated from a mixture of $N(0, 0.03^2)$, $N(0.5, 0.03^2)$ and $N(1, 0.03^2)$. Similarly, for the other copy number states with multiple possible genotypes, we randomly generated the genotype for each location.

We compared the performance of modSaRa2 to existing state-of-the-art methods, namely, CBS (Venkatraman and Olshen, 2007) and PennCNV (Wang *et al.*, 2007). We also focused on the improved sensitivity and specificity as compared the original method, modSaRa, to demonstrate the characteristics of the new method.

2.6 Application dataset and quality control steps

To illustrate the proposed method, we further applied modSaRa2 to a cohort study of melanoma from Gene Environment Association Studies initiative (GENEVA) that included 3115 participants. The high-density SNP array data for skin cutaneous melanoma were released in 2008, details of which have been described previously (Amos *et al.*, 2011). The raw data were processed by Illumina's genotyping module v1.94 in the GenomeStudio (v2.0.2) to calculate probe intensities including LRR and BAF. Samples with a standard deviation of LRR on chromosome 1 less than 0.25 were high quality intensity data for CNV analysis and therefore retained. Missing data, presence of more homozygous genotypes or departure from Hardy Weinberg Equilibrium (HWE) in samples can be possibly caused by duplications or deletions on chromosomal segments. Thus, typical quality control (QC) procedures in genotyping based on HWE deviation or low call rate of SNPs for CNV analysis were not employed in this study. QC filters in this study only included exclusion of duplicates or relatives based on pairwise identity by descent calculation (IBD > 0.95). For those pairs of duplicates, samples with relatively higher call rate were retained.

2.7 CNV calling and association with melanoma risk

After intensities pre-processing step of GC model adjustment by PennCNV, modSaRa2 was applied in CNV calling. After obtaining the generated raw CNVs, we merged adjacent CNVs with distance less than ten markers. CNV quality control filters included retaining samples with a total number of CNVs < 1000; CNVs with length $>$ 10 markers, > 10 kb and < 1 Mb. NCBI build 36 (hg18) was used for finding overlapping genes for CNV calls. The identified CNVs were mapped to 1000 Genome Project phase 3 which curated information about 60 000 structural variations captured at the population level (Sudmant *et al.*, 2015), and the database for human CNV map (Zarrei *et al.*, 2015) to ascertain CNV calls on 935 medically relevant genes.

After CNV calling, a gene-based association strategy was applied to investigate the contribution of CNVs to melanoma risk susceptibility. The association of the CNVs in each gene with melanoma risk was evaluated using logistic regression $\log it(P(Y=1)) = \beta_0 + \beta_1 I_{del} + \beta_2 I_{dup} + \beta_3 Age + \beta_4 Gender$. Effects from deletions and duplications were tested separately; gender and age were adjusted as covariates. Permutation based adjustment was performed so that 10 000 replicates in each test were generated to adjust the nominal *P*-values.

3 Results

3.1 Performance assessment via simulations

We assessed the performance of modSaRa2 in different scenarios and compared it to modSaRa and two other conventional methods, PennCNV and CBS (Table 1 and Fig. 1). Table 1 provides an overall comparison for all simulation scenarios with different CNV sizes and copy number states. The Receive Operating Characteristic (ROC) curve in Figure 1 explicitly compared these methods in detecting weak signals at various threshold setting (Dup.S). Overall,

all four tests increased power when the jump sizes increased. For gain/loss of two copies (Del.D and Dup.D), all methods detected almost all breakpoints accurately. The major differences were in detecting weak signals (Del.S and Dup.S) that modSaRa2 easily achieved a high level of true positive rate (TPR) while controlling the type I error very well (Fig. 1), details of which are illustrated below.

First, we demonstrated the increased sensitivity and specificity of modSaRa2 compared to the modSaRa in detecting weak signals (Dup.S). At a comparable or higher TPR, modSaRa2 identified much fewer false positives (FPs) than modSaRa. For short CNVs with (length 10–50), with a higher TPR at around 0.99, modSaRa2 detected fewer FPs than modSaRa (0 versus 12). When the length of CNV segments increase, the number of FPs remained at a low level for modSaRa2 whereas the number detected by modSaRa increased. A receiver operating characteristic (ROC) curve further compared these two methods explicitly (Fig. 1). In summary, modSaRa2 was improved in specificity compared to the original modSaRa method for detecting CNVs with weak signals.

To evaluate the performance of change point model-based and Hidden Markov model-based methods, the second comparison was between the modSaRa methods (i.e. modSaRa2 and modSaRa) and PennCNV. Still, there were no differences in detecting loss/gain of double copies (Del.D and Dup.D) (Table 1). For detection of Dup.S, modSaRa2 and modSaRa outperformed PennCNV in most scenarios of weak signals. For detection of Del.S, modSaRa2 outperformed PennCNV when CNVs were relatively large in length (>50 markers). modSaRa2 showed consistency in performance; however, PennCNV was less powerful when CNV sizes increased.

Although the comparison of modSaRa and CBS has been thoroughly illustrated (Xiao *et al.*, 2015), we compared modSaRa2 and CBS to be inclusive. In almost all scenarios, modSaRa2 outperformed CBS for detection of weak signals (Fig. 1). For detection of Del.D with small sizes, modSaRa2 achieved higher specificity than CBS. In conclusion, modSaRa2 showed superior performance in detecting CNVs, and its accuracy in detecting weak CNV signals was significantly improved by integrating more external genetic information. It is worth mentioning that CBS was the slowest algorithm among the four tests in our study.

Table 1. Assessment of power and false discovery rate with different CNV sizes and jump sizes

		modSaRa2		modSaRa		PennCNV		CBS	
CNV length	CNV State	TPR	#FP	TPR	#FP	TPR	#FP	TPR	#FP
10–50	Del.D	1.0	0	1.0	0	1.0	0	1.0	23
	Del.S	0.9990	0	1.0	0	1.0	0	0.9990	3
	Dup.S	0.9860	0	0.8890	12	0.9500	0	0.9460	13
	Dup.D	1.0	0	1.0	0	1.0	0	1.0	1
50–100	Del.D	1.0	0	1.0	0	1.0	0	1.0	1
	Del.S	1.0	0	1.0	0	0.95	100	1.0	0
	Dup.S	0.9870	0	0.9755	23	0.9750	50	0.9910	17
	Dup.D	1.0	0	1.0	0	1.0	0	1.0	0
100–200	Del.D	1.0	0	1.0	0	1.0	0	1.0	0
	Del.S	0.9995	1	0.9995	1	0.9000	100	0.9995	3
	Dup.S	0.9910	0	0.9695	51	0.9250	150	0.9865	36
	Dup.D	1.0	0	1.0	0	1.0	0	1.0	0

Del.D, deletion of double copy; Del.S, deletion of single copy; Dup.S, duplication of single copy; Dup.D, duplication of double copy. The total number of simulated CNV was 1, 000. Significance level of modSaRa2 was set as the default value of 0.01. TPR, True positive rate; #FP, Number of false positives.

3.2 Analysis of whole genome data of melanoma cases and controls

In this study, we identified 354 210 CNVs in autosomes of 2838 samples (see Fig. 2 for study design). Overall, the total number of deletions was nearly as large as duplications (Supplementary Table S3). No significant difference was observed in the overall proportion of CNVs with deletions in cases and controls, 48% versus 46%, respectively. Also, there was no difference in the number or length of CNVs between cases and controls.

Different CNV calling methods have their strengths and limitations (Kumaran *et al.*, 2017; Sapkota *et al.*, 2016); the CNV break points called by different algorithms may or may not overlap and some algorithms tend to call redundant CNVs. Therefore, it was important to ascertain that the called CNVs were reliable by independent methods. We therefore considered higher resolution structural variation data available from the public domain from the 1000 Genomes Project (Phase 3) as a reference (Sudmant *et al.*, 2015). CNVs were mapped to 1000 Genomes Project data to access concordances for the CNVs identified in this study (Supplementary Table S4). Of the break points identified by modSaRa2, 28.49% were mapped to the 1000 Genomes Project, which was slightly higher than those identified by PennCNV (26.63%). Another comparison was the empirical true positive rate. We mapped 97.12% of the clustered common copy number region from the 1000 Genomes Project to the breakpoints called by modSaRa2, which was much higher

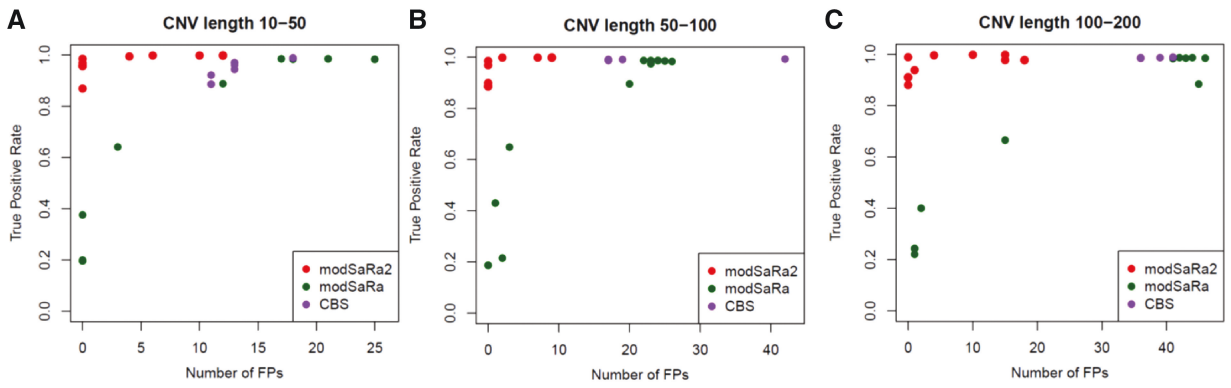

Fig. 1. Evaluation of statistical power through simulations. Detailed simulation settings are described under Section 2. Results of scenario of duplication of single copy (Dup.S) with (**A**) short CNV sizes (10–50 markers); (**B**) medium CNV sizes (50–100 markers); (**C**) long CNV sizes (100–200 markers)

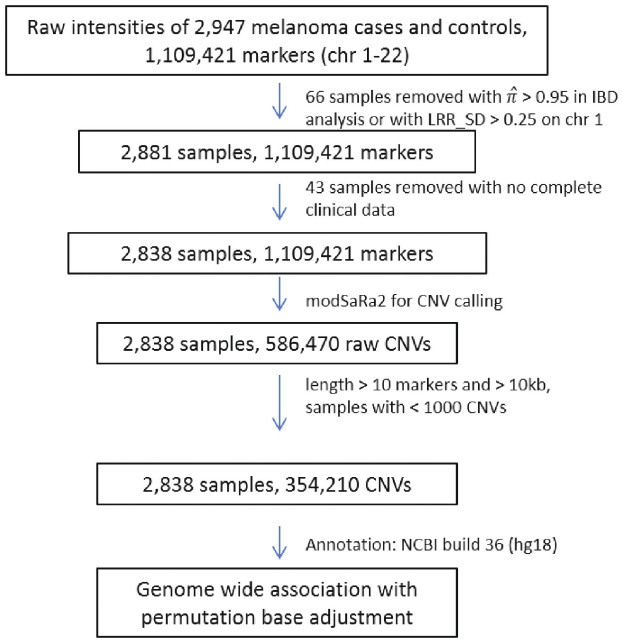

Fig. 2. Overview of the whole genome CNV study. The figure outlines the study design with brief description of data filters and methods. Summary of key results of each analysis indicating the sample size and number of CNVs at various stage of analysis

than PennCNV (51.09%). This comparison was not ideal as the results from 1000 Genomes Project cannot be considered as truth, but it still provides reference about the high sensitivity of modSaRa2.

Although we identified CNVs in both protein coding and non-protein coding regions, those overlapping protein coding regions have higher potential to contribute to variation in the trait (Lee and Scherer, 2010). Therefore, CNVs overlapping with protein coding genes were considered for melanoma risk relevance. Coding genes within the CNVs identified were interrogated for gene-dosage effects by evaluating the association between copy number status and melanoma status. We identified 133 genes within the deletions significantly associated with melanoma (permutation P-value < 0.05). Of these, 34 genes were overlapping with the 1000 Genomes Project curated structural variations or human CNV map (Table 2, Supplementary Table S6). We also identified 161 genes with duplications that showed significant association with melanoma risk (Supplementary Table S7), among which 41 were overlapping with the 1000 Genomes Project data or human CNV map. Among the significantly associated deletions were gene *LRRC8C* in 1p22.2 region (chr1: 89 842 228–89 893 620) (OR = 2.11, 95% CI = 1.12–4.32, adjusted P-values = 2.35×10^{-2}, Supplementary Table S6). A plot of the LRR and BAF indicated that all of these variants were deletions of single copy (data not shown). The amplification of the same region was not statistically significant (P-value = 0.35). The results of duplication highlighted the chromosomal regions on 6p22,

Table 2. Top significantly associated germline CNVs with melanoma risk (deletions) with mapping to publicly available databases (P-value < 0.001)

			Deletions			Duplications		
Gene	Chr	Coordinate	Case:Cont	OR (95%CI)	P.adj	OR (95%CI)	P.adj	Mapping
LIFR	5p13.1	38 510 821–38 631 264	98:18	2.98 (1.84, 5.12)	1.00×10^{-4}	0.63 (0.47, 0.86)	4.80×10^{-3}	–
ZFYVE9	1p32.3	52 380 353–52 584 946	95:18	2.86 (1.76, 4.91)	1.00×10^{-4}	0.52 (0.37, 0.72)	2.00×10^{-4}	–
MMRN2	10q23.2	88 685 277–88 707 405	63:11	3.10 (1.69, 6.24)	2.00×10^{-4}	0.54 (0.35, 0.83)	3.30×10^{-3}	–
ANKRD33B	5p15.2	10 617 434–10 710 928	105:27	2.15 (1.42, 3.38)	5.00×10^{-4}	0.91 (0.73, 1.13)	0.39	–
LIFR-AS1	5p13.1	38 592 644–38 644 716	77:17	2.47 (1.49, 4.34)	5.00×10^{-4}	0.63 (0.45, 0.89)	7.10×10^{-3}	1000g
PL × ND1	3q22.1	130 756 745–130 808 272	54:10	2.93 (1.55, 6.14)	7.00×10^{-4}	0.67 (0.42, 1.07)	8.52×10^{-2}	–
SNCG	10q23.2	88 708 267–88 712 997	33:2	9.23 (2.79, 57.07)	9.00×10^{-4}	0.37 (0.20, 0.66)	3.00×10^{-4}	–
MMP15	16q21	56 616 782–56 638 305	69:16	2.37 (1.40, 4.25)	1.30×10^{-3}	0.98 (0.68, 1.44)	0.93	–
CLSTN2	3q23	141 136 716–141 769 609	64:14	2.47 (1.42, 4.61)	2.20×10^{-3}	0.67 (0.47, 0.94)	1.61×10^{-2}	1000g
MARCH11	5p15.1	16 120 473–16 232 897	124:39	1.74 (1.21, 2.54)	2.60×10^{-3}	0.74 (0.53, 1.02)	6.85×10^{-2}	–
MIR4656	7p22.1	4 794 721–4 794 796	86:21	2.07 (1.30, 3.44)	2.60×10^{-3}	0.47 (0.37, 0.59)	1.00	–
HTRA3	4p16.1	8 322 388–8 359 738	32:3	5.79 (2.07, 24.13)	2.80×10^{-3}	0.47 (0.26, 0.84)	9.60×10^{-3}	–
ITGB5	3q21.2	125 964 484–126 088 834	29:2	7.87 (2.37, 48.76)	2.90×10^{-3}	0.63 (0.36, 1.12)	0.13	–
AC131097.3	2q37.3	242 472 186–242 669 546	214:81	1.48 (1.14, 1.94)	3.70×10^{-3}	1.31 (0.97, 1.79)	9.11×10^{-2}	–
GABRA5	15q12	24 663 365–24 777 103	99:30	1.80 (1.20, 2.77)	5.10×10^{-3}	0.63 (0.46, 0.88)	7.00×10^{-3}	–
MYO5B	18q21.1	45 603 153–45 975 449	129:44	1.64 (1.16, 2.36)	5.20×10^{-3}	1.28 (0.89, 1.85)	0.19	–
CCDC85C	14q32.2	99 047 355–99 140 480	19:1	10.51 (2.17, 189.18)	5.30×10^{-3}	0.52 (0.35, 0.78)	1.70×10^{-3}	1000g
LYN × 1	8q24.3	143 842 757–143 855 746	22:2	6.14 (1.80, 38.49)	5.60×10^{-3}	0.48 (0.22, 1.03)	4.21×10^{-2}	–
RNU6-71P	13q13.3	56 245 647–56 723 097	63:16	2.17 (1.28, 3.90)	5.90×10^{-3}	0.26 (0.01, 2.73)	6.18×10^{-2}	1000g
DENND5BAS1	12p11.21	31 634 123–31 659 552	38:7	2.90 (1.37, 7.12)	7.00×10^{-3}	0.51 (0.32, 0.81)	4.00×10^{-3}	–
DENND5B	12p11.21	31 426 423–31 635 219	38:7	2.90 (1.37, 7.11)	8.20×10^{-3}	0.49 (0.31, 0.79)	1.90×10^{-3}	–
NR2F6	19p13.11	17 203 693–17 217 151	31:5	3.39 (1.43, 9.95)	8.60×10^{-3}	0.48 (0.24, 0.94)	2.54×10^{-2}	–
USHBP1	19p13.11	17 221 848–17 236 544	31:5	3.39 (1.43, 9.95)	8.60×10^{-3}	0.45 (0.22, 0.89)	1.66×10^{-2}	–
AP5Z1	7p22.1	4 781 787–4 800 552	100:28	1.79 (1.18, 2.79)	9.00×10^{-3}	0.48 (0.38, 0.60)	1.00	–
RGS12	4p16.3	3 285 671–3 411 438	37:7	2.89 (1.37, 7.11)	9.00×10^{-3}	1.29 (0.77, 2.23)	0.32	1000g

P.adj is the adjusted P-value after 10 000 permutations. chr, chromosome; 1000g, 1000 Genomes Project; CNV map, the curated medically relevant CNVs. The column of Case, Cont provides the number of CNVs in cases and controls, respectively. To illustrates the risk variants associated with melanoma, we provided results of variants with OR greater than 1.

6q25 and 19p13 (Supplementary Table S5). Still, the deletion of any of these regions was not statistically significant. It was noteworthy that most of the identified candidate associated CNVs were all rare variants, with sample proportions ranging from 1% to 2%, which may partially explain why the discoveries of germline CNVs in cancer studies are still not compelling.

4 Discussion

This paper mainly is directed at large-scale CNV detection for improving specificity of CNVs with moderate jump sizes which remains the bottleneck for achieving high accuracy in CNV detection. Massive genome-wide data for associating germline variants with diseases and traits have been produced in the past few decades, but the important roles of germline CNVs have not been fully explored. To date, the majority of germline CNVs that have been identified for diseases are either rare conferring high penetrance, or common variants with low penetrance, especially in cancer (Al-Sukhni *et al.*, 2012; Krepischi *et al.*, 2012; Kuusisto *et al.*, 2013; Laitinen *et al.*, 2016). To identify statistically significant DNA copy number changes, proper statistical modeling is critical, especially when some CNV signals are difficult to capture.

A limitation shared by many compelling CNV detection methods that have been highlighted by multiple independent benchmarking studies is the lack of sensitivity and specificity for variants with weak CNV signals. To meet the widespread demand for improved CNV detection, we developed a new method, modSaRa2, to reduce the technical noise by integrating more genetic information and external empirical statistics in statistical modeling. modSaRa2 builds on our existing method modSaRa with the significant improvement of sensitivity and specificity (Supplementary Figs. S2–S3), thus allowing full-spectrum CNV detection. modSaRa2 can be applied to profile all CNVs in multiple platforms of microarray data and has the potential to be extended to next generation sequencing generated signals.

We thoroughly evaluated modSaRa2 against existing methods. First, we performed extensive simulations to elucidate how key variables, such as CNV sizes and jump sizes, influence performance in detection. We showed that modSaRa2 markedly improved both sensitivity and specificity over existing methods. The improvement for weak CNV signals was the most substantial, with simultaneously improved stability when CNV size varied. The newly computed signal intensity that integrated both LRR and BAF coupled with the distribution assumption with prior knowledge of the intensities enhanced the signals from CNVs with small jump sizes. In the second evaluation, we used an ROC curve to demonstrate the remarkably improved performance of modSaRa2 versus modSaRa. Finally, we applied modSaRa2 to whole genome microarray data of cutaneous melanoma, where modSaRa2 detected CNVs with higher sensitivity and comparable precision than PennCNV. In combination, these results established the improved accuracy of modSaRa2 over other state-of-the-art approaches, as well as the stability under various conditions.

modSaRa2 is the first attempt to use change point searching method in a comprehensive framework by fully utilizing genetic intensities and empirical statistics of the intensities. It uses the empirical values from the Illumina platform generated data and allows customized usage by inputting empirical statistics of other platforms (e.g. Affymetrix). With the increasing capacity of CNV information and a rising need to profile CNVs as an essential source of genetic variation, modSaRa2 is a flexible approach that has the potential to be widely used. We are exploring ways to extend the proposed method to analyze sequencing data. As we know, next generation sequencing data present different features from array data, including discrete nature and higher dimensionality, which sets untrivial barriers for the natural extension from methods originally developed for microarray. With existing sequencing data-based CNV analysis tools, continuous intensities can be generated after normalization of read counts and comparison to a reference panel (Jiang *et al.*, 2015; Magi *et al.*, 2013). Segmentation methods such as CBS are then applied for breakpoints identification (Jiang *et al.*, 2015). We conducted preliminary studies with these ideas, and it appeared it is feasible to extend our method to the sequence data, but it is beyond the scope of this work to present our extension.

We have presented the results of applying modSaRa2 to whole genome germline CNV calling and the association with melanoma risk. Among the top associations is the discovery of 1p22.2 region overlapping with *LRRC8C* that was previously found in a region containing a novel susceptibility gene for cutaneous melanoma by linkage analysis (Gillanders *et al.*, 2003). Moreover, deletion mapping with somatic melanoma tissues or cell lines suggested that a tumor suppressor gene was contained in this region (Walker *et al.*, 2004). Copy number aberration analysis with TCGA samples also indicated that the deletion of this region was significantly associated with melanoma tumorigenesis (http://gdac.broadinstitute.org). In addition, the amplification of chromosome 6p has been described, and the chromosome 6q25 was found to be frequently altered in diverse tumor types including melanoma (Millikin *et al.*, 1991; Santos *et al.*, 2007). Our above finding indicated the possible roles of deletions on 1p22.2 region and the amplifications on 6p22 and 6q25 in the origin and tumorigenesis of melanoma, although further validation studies need to be performed.

Acknowledgements

We acknowledge Gene Environment Association Studies initiative (GENEVA) for providing the melanoma dataset for our study.

Funding

This work was supported by the National Science Foundation (NSF) grant DMS1722562 (Xiao), NSF grant DMS 1722691 (Niu and Hao), NSF grant DMS1722544 (Zhang), National Institute of Health (NIH) grant R01DA016750 (Zhang), NIH grant R01MH116527 (Zhang), and the internal ASPIRE-I fund from the University of South Carolina (Xiao).

Conflict of Interest: none declared.

References

Al-Sukhni,W. *et al.* (2012) Identification of germline genomic copy number variation in familial pancreatic cancer. *Hum. Genet.*, **131**, 1481–1494.

Amos,C.I. *et al.* (2011) Genome-wide association study identifies novel loci predisposing to cutaneous melanoma. *Hum. Mol. Genet.*, **20**, 5012–5023.

Barnes,C. *et al.* (2008) A robust statistical method for case-control association testing with copy number variation. *Nat. Genet.*, **40**, 1245–1252.

Chen,M. *et al.* (2013a) SomatiCA: identifying, characterizing and quantifying somatic copy number aberrations from cancer genome sequencing data. *PLoS One*, **8**, e78143.

Chen,W. *et al.* (2013b) Identification of chromosomal copy number variations and novel candidate loci in hereditary nonpolyposis colorectal cancer with mismatch repair proficiency. *Genomics*, **102**, 27–34.

Conrad,D.F. *et al.* (2010) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**, 704–712.

Darvishi,K. (2010) Application of Nexus copy number software for CNV detection and analysis. *Curr. Protoc. Hum. Genet.*, **4**, 1–28.

Dempster,A.P. *et al*. (1977) Maximum likelihood from incomplete data via EM algorithm. *J. R. Stat. Soc. B Met*., **39**, 1–38.

Deng,X. (2011) SeqGene: a comprehensive software solution for mining exome- and transcriptome-sequencing data. *BMC Bioinformatics*, **12**, 267.

Gai,X. *et al*. (2010) CNV Workshop: an integrated platform for high-throughput copy number variation discovery and clinical diagnostics. *BMC Bioinformatics*, **11**, 74.

Gillanders,E. *et al*. (2003) Localization of a novel melanoma susceptibility locus to 1p22. *Am. J. Hum. Genet*., **73**, 301–313.

Hollox,E.J. *et al*. (2008) Psoriasis is associated with increased beta-defensin genomic copy number. *Nat. Genet*., **40**, 23–25.

Hou,S. *et al*. (2013) Copy number variations of complement component C4 are associated with Behcet's disease but not with ankylosing spondylitis associated with acute anterior uveitis. *Arthritis Rheum*., **65**, 2963–2970.

International HapMap, C. *et al*. (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–58.

Jiang,Y. *et al*. (2015) CODEX: a normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Res*., **43**, e39.

Krepischi,A.C. *et al*. (2012) Germline DNA copy number variation in familial and early-onset breast cancer. *Breast Cancer. Res*., **14**, R24.

Kumaran,M. *et al*. (2017) Germline copy number variations are associated with breast cancer risk and prognosis. *Sci. Rep*., **7**, 14621.

Kuusisto,K.M. *et al*. (2013) copy number variation analysis in familial BRCA1/2-negative Finnish breast and ovarian cancer. *PLoS One*, **8**, e71802.

Laitinen,V.H. *et al*. (2016) Germline copy number variation analysis in Finnish families with hereditary prostate cancer. *Prostate*, **76**, 316–324.

Lee,C. and Scherer,S.W. (2010) The clinical context of copy number variation in the human genome. *Expert Rev. Mol. Med*., **12**, e8.

Li,N. *et al*. (2017) Association between C4, C4A, and C4B copy number variations and susceptibility to autoimmune diseases: a meta-analysis. *Sci. Rep*., **7**, 42628.

Lin,C.H. *et al*. (2011) Molecular profile and copy number analysis of sporadic colorectal cancer in Taiwan. *J. Biomed. Sci*., **18**, 36.

Magi,A. *et al*. (2013) EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biol*., **14**, R120.

Maher,B. (2008) Personal genomes: the case of the missing heritability. *Nature*, **456**, 18–21.

Marshall,C.R. *et al*. (2017) Contribution of copy number variants to schizophrenia from a genome-wide study of 41, 321 subjects. *Nat. Genet*., **49**, 27–35.

McCarroll,S.A. *et al*. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet*., **40**, 1166–1174.

Millikin,D. *et al*. (1991) Loss of heterozygosity for loci on the long arm of chromosome 6 in human malignant melanoma. *Cancer Res*., **51**, 5449–5453.

Niu,Y.S. and Zhang,H. (2012) The screening and ranking algorithm to detect DNA copy number variations. *Ann. Appl. Stat*., **6**, 1306–1326.

Olshen,A.B. *et al*. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Peiffer,D.A. *et al*. (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res*., **16**, 1136–1148.

Qiu,Z.W. *et al*. (2017) Genome-wide copy number variation pattern analysis and a classification signature for non-small cell lung cancer. *Genes Chromosomes Cancer*, **56**, 559–569.

Santos,G.C. *et al*. (2007) Chromosome 6p amplification and cancer progression. *J. Clin. Pathol*., **60**, 1–7.

Sapkota,Y. *et al*. (2016) A genome-wide association study to identify potential germline copy number variants for sporadic breast cancer susceptibility. *Cytogenet. Genome Res*., **149**, 156–164.

Stuart,P.E. *et al*. (2012) Association of beta-defensin copy number and psoriasis in three cohorts of European origin. *J. Invest. Dermatol*., **132**, 2407–2413.

Sudmant,P.H. *et al*. (2015) An integrated map of structural variation in 2, 504 human genomes. *Nature*, **526**, 75–81.

Venkatraman,E.S. and Olshen,A.B. (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, **23**, 657–663.

Walker,G.J. *et al*. (2004) Deletion mapping suggests that the 1p22 melanoma susceptibility gene is a tumor suppressor localized to a 9-Mb interval. *Gene Chromosome Cancer*, **41**, 56–64.

Wang,K. *et al*. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res*., **17**, 1665–1674.

Xiao,F. *et al*. (2015) Modified screening and ranking algorithm for copy number variation detection. *Bioinformatics*, **31**, 1341–1348.

Xiao,F. *et al*. (2017) modSaRa: a computationally efficient R package for CNV identification. *Bioinformatics*, **33**, 2384–2385.

Zarrei,M. *et al*. (2015) A copy number variation map of the human genome. *Nat. Rev. Genet*., **16**, 172–183.

Zhang,N.R. (2010) DNA copy number profiling in normal and tumor genomes. *Comput. Biol. Ser*., **15**, 259–281.