

Summer 2021

Accurate and Integrative Detection of Copy Number Variants With High-Throughput Data

Xizhi Luo

Follow this and additional works at: <https://scholarcommons.sc.edu/etd>



Part of the [Biostatistics Commons](#)

Recommended Citation

Luo, X.(2021). *Accurate and Integrative Detection of Copy Number Variants With High-Throughput Data*. (Doctoral dissertation). Retrieved from <https://scholarcommons.sc.edu/etd/6413>

This Open Access Dissertation is brought to you by Scholar Commons. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact digres@mailbox.sc.edu.

ACCURATE AND INTEGRATIVE DETECTION OF COPY NUMBER VARIANTS
WITH HIGH-THROUGHPUT DATA

by

Xizhi Luo

Bachelor of Medicine
Guangzhou Medical University, 2009

Master of Science
Rutgers University, 2016

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Biostatistics

Department of Epidemiology & Biostatistics

Arnold School of Public Health

University of South Carolina

2021

Accepted by:

Feifei Xiao, Major Professor

Bo Cai, Committee Member

Alexander C. McLain, Committee Member

Guoshuai Cai, Committee Member

Tracey L. Weldon, Interim Vice Provost and Dean of the Graduate School

© Copyright by Xizhi Luo, 2021
All Rights Reserved.

ACKNOWLEDGEMENTS

It is a great honor to express my gratitude and sincerest appreciation to my advisor Dr. Feifei Xiao, who is an outstanding and enthusiastic researcher and a great mentor. Her mentoring and encouragement have been solely and mainly responsible for completing everything in this dissertation. My sincerest thanks also go to Dr. Guoshuai Cai. It is my luck to be advised by both Dr. Xiao and Dr. Cai, from whom I learnt not only how to be a qualified Ph.D. student but also how to be a thinker. Thank you for sharing your valuable insights and experiences, and for always being supportive and patient in letting me pursue my research interest.

I am also grateful to my dissertation committee members, Dr. Bo Cai, and Dr. Alexander C McLain, who are successful statistician and inspiring educator. Thanks for providing invaluable advice and suggestions in guiding my research and overcoming difficulties.

My special thanks also go to Ph.D. students in the Biostatistics department: Yuan Hong, Yanan Zhang, Fei Qin, and Xuanxuan Yu. Thank you all for making my PhD life and day colorful.

Finally, it is my privilege to thank my parents and my wife Mrs. Zijie Shen, who do not care about my titles and achievements, but only want me to be a better myself and live a fulfilling life. Thanks for being beside me all the time.

ABSTRACT

Copy number variation, as a major source of genetic variation in the human genome, are gains or losses of the DNA segments. Copy number variation has gained considerable interest as it plays important roles in human complex diseases. Therefore, accurate detection of CNVs with data generated by modern genotyping technologies, such as SNP array and whole-exome sequencing (WES), comprises a critical step toward a better understanding of disease etiology. However, current statistical methodologies for CNV detection still face analytical challenges due to numerous genetic and technological factors that may lead to spurious findings. First, existing methods assume the independent observations along the whole genome in genetic intensities for CNV detection, which is often violated in the genetics perspective. Second, neither SNP array nor WES offers full coverage of the genome in their genotyping resolution, which leads to a significant amount of missed variant calls by analyzing each data separately. Third, conventional methods adopt a single sample-based strategy that suffers from high false discovery rates due to prominent data noise.

In this study, we developed (a) a SNP array CNV calling algorithm, LDcnv, that integrated the genomic correlation structure with a local search strategy into statistical modelling of the genetic intensities, which improves both detection accuracy and robustness; (2) a WES CNV detection method, CORRseq, that extended the methodological work of LDcnv coupled with a median normalization procedure, which

gives significant power gain for CNV identification; (3) a Bayesian Multi-sample and Integrative CNV (BMI-CNV) profiling method with matched samples sequenced by both WES and microarray, which used a Bayesian probit stick-breaking process model coupled with a Gaussian Mixture Model estimation for multiple sample integration. BMI-CNV enables accurate CNV identification integrating multiple genotyping platforms with a genome-wide scale. The performance of these proposed methods has been evaluated by extensive simulation studies and real data analyses. Our novel methods have been further applied to the 1000 Genomes project and an international lung cancer study to identify lung cancer susceptibility genes.

The proposed framework has a broad application scope for multiple study designs in studying the role of CNVs in various complex diseases, which will reveal the vital roles of CNVs in disease development and inspire new approaches for precision medicine.

TABLE OF CONTENTS

Abstract	iv
List of Tables	viii
List of Figures	x
Chapter 1 INTRODUCTION	1
1.1 COPY NUMBER VARIATION AND HUMAN DISEASES	1
1.2 STATISTICAL METHODS FOR CNV DETECTION	2
CHAPTER 2 INTEGRATING GENOMIC CORRELATION STRUCTURE IMPROVES COPY NUMBER VARIANTS DETECTION	11
2.1 ABSTRACT	11
2.2 INTRODUCTION	11
2.3 METHODS	13
2.4 RESULTS	20
2.5 DISCUSSION	25
CHAPTER 3 INTEGRATING GENOMIC CORRELATION STRUCTURE ENHANCES COPY NUMBER VARIANTS DETECTION FROM WHOLE-EXOME SEQUENCING DATA	31
3.1 ABSTRACT	31
3.2 INTRODUCTION	32
3.3 METHODS	35
3.4 RESULTS	40
3.5 DISCUSSION	45

CHAPTER 4 BMI-CNV: A BAYESIAN FRAMEWORK FOR MULTIPLE GENOTYPING PLATFORMS DETECTION OF COPY NUMBER VARIANTS	56
4.1 ABSTRACT	56
4.2 INTRODUCTION	57
4.3 METHODS	59
4.4 RESULTS	68
4.5 DISCUSSION	71
CHAPTER 5 CONCLUSIONS AND FUTURE PERSPECTIVES	81
References	86
Appendix A: Supplementary figures and tables	94
Appendix B: Integrating genomic correlation structure improves copy number variants detection	112
Appendix C: BMI-CNV: a Bayesian framework for multiple genotyping platforms detection of copy number variants	117
Appendix D: Permission to reprint	127

LIST OF TABLES

Table 2.1 Summary of CNV calls on simulated data at $\phi = 0.3$ from all methods.	25
Table 3.1 Assessment of CORRseq and SARAsq based on simulation data.....	44
Table 3.2 Assessment of CORRseq and SARAsq based on the 1000 Genomes Project data.....	45
Table 4.1 Summary of performance of our method on simulated data in the integrative analysis.	70
Table 4.2 Summary of performance of our method on simulated data in WES analysis.	71
Table A.1 Joint genotype probabilities for two diallelic loci.....	94
Table A.2 Relationship between CNV locations and LD map	95
Table A.3 Summary of CNV calls on simulated data at $\phi = 0.1$ from all methods	96
Table A.4 Summary of CNV calls on simulated data at $\phi = 0.5$ from all methods	97
Table A.5 Overall assessment of CNV calling on the HapMap project dataset	98
Table A.6 Assessment of calling performance in short CNVs on the HapMap project dataset.	99
Table A.7 Demographic characteristics of the study participants after quality control filters.....	100
Table A.8 Summary of joint CORRseq and EXCAVATOR2 calling results on TRICL data.	101
Table A.9 Top significantly associated CNVs with lung cancer risk (P-value<0.01)	102
Table A.10 Top significantly associated CNVRs with lung cancer risk (P-value<0.01)	103

Table A.11 Summary of BMI-CNV calling results on TRICL data.....	104
Table A.12 Top significantly associated CNVs with lung cancer risk (P-value<0.005)	105
Table B.1 All possible Bivariate distributions and assigned Genotypes.	113
Table C.1 Copy number states and empirical parameter settings for SNP array data.	126
Table C.2 Copy number states and empirical parameter settings for WES data.	126

LIST OF FIGURES

Figure 2.1 Assessment of CNV calls generated by LDcnv, PennCNV, CBS and SLMSuite methods with validation datasets from (a) HapMap 3 (b) Conrad et al (c) McCarroll et al.	26
Figure 3.1 Analysis workflow of CORRseq and SARAsq	46
Figure 3.2 Assessment of genomic correlation structure of WES data from 1000 Genomes project	47
Figure 3.3 Assessment of CORRseq and SARAsq based on simulation data	48
Figure 3.4 Assessment of CORRseq and SARAsq based on the 1000 Genomes Project data.....	49
Figure 3.5 Overview of the application to the TRICL case-control study.....	50
Figure 3.6 CNVs identified by CORRseq from the TRICL study.....	51
Figure 4.1 Analysis workflow of BMI-CNV	72
Figure 4.2 Performance assessment of BMI-CNV and iCNV on simulated data in the integrative analysis	73
Figure 4.3 Performance assessment of BMI-CNV, iCNV, EXCAVATOR2, and CODEX2 on simulated data in WES analysis	74
Figure 4.4 Overview of the application to the 1000 genomes project and HapMap data	75
Figure 4.5 Overview of the integrative analysis of the TRICL case-control study	76
Figure A.1 Four classifications of the CNV locations in the LD genome map	106
Figure A.2 Assessment of CNV calling performance in short CNVs on the HapMap project datasets	107
Figure A.3 Summary of length and frequency of BMI-CNV calling results on 1000 genome and HapMap project data.....	108

Figure A.4 Case illustration of one common deletion region of chromosome 3	109
Figure A.5 Comparison of length and frequency of BMI-CNV calling results on 1000 genome and HapMap project data under two different data integration strategies.....	110
Figure A.6 Data intensity of top lung cancer related CNVs	111

CHAPTER 1

INTRODUCTION

1.1 COPY NUMBER VARIATION AND HUMAN DISEASES

Human genome varies considerably from one to another in numerous ways. Those genetic variations form the heritability of complex human traits. Progress has been made in revealing millions of single nucleotide polymorphisms (SNPs) as the predominant form of genomic variation (A *et al.*, 2019). However, copy number variation that constitutes a large category of structural variation, involving genomic deletions and duplications, is still not well-studied. In 2004, two research groups reported the widespread presence of copy number variations in normal individuals and their role as a significant source of benign genomic variation (Zhang *et al.*, 2009). Since then, many studies have been conducted, and over 202,431 copy number variants (CNVs) have now been reported (MacDonald *et al.*, 2014).

To understand the mechanisms of copy number variation formation, two well-recognized recombination-based theories have been implicated: nonallelic homologous recombination (NAHR) and nonhomologous end-joining (NHEJ) (Lupski, 1998; Lupski and Stankiewicz, 2002; Schwarz *et al.*, 2003; Lieber, 2008). NAHR occurs when two DNA sequences share high sequence similarities. During meiosis or mitosis, low copy repeats (LCRs) which are substrates for NAHR can be misaligned, leading to genetic rearrangement. Deletions or duplications occur when NAHR occurs between different

LCRs. On the other hand, NHEJ is mainly responsible for repairing DNA double-strand breaks. This error-prone process can create loss or addition of several nucleotides at the joining point (Lupski, 1998; Lupski and Stankiewicz, 2002; Rodgers and Mcvey, 2016). With the use of the array-based techniques, many NAHR or NHEJ-mediated genomic deletions and duplications have been observed in multiple studies (Schwarz *et al.*, 2003; Lieber, 2008).

In addition to generating polymorphism in the population, CNVs have also been shown to be associated with susceptibility to diseases. For example, CNVs have been extensively studied and demonstrated as genetic determinants for lung cancer, familial breast cancer, and melanoma (Frank *et al.*, 2007; Lesueur *et al.*, 2008; Qiu *et al.*, 2017). Sebat *et al.* identified a deletion of 16q11.2 and duplication of 15q11-13 as autism susceptibility loci (J *et al.*, 2007). Fellermann *et al.* showed that colonic Crohn disease patients had significantly fewer copies of beta-defensin gene (HBD-2) than healthy controls, and they demonstrated that individuals with three or fewer copies of HBD-2 had a higher risk of developing Crohn disease (Fellermann *et al.*, 2003, 2006). Also, a recurrent deletion at 17q12 has been revealed to be associated with an increased risk of ASD and schizophrenia (Moreno-De-Luca *et al.*, 2010). A common CNV in the *Hp* gene has been discovered to be a risk factor for cardiovascular disease in patients with type 2 diabetes (Wang *et al.*, 2019).

1.2 STATISTICAL METHODS FOR CNV DETECTION

Currently, comprehensive profiling of CNVs commonly relies on two major genome-wide genotyping technologies: SNP array and next-generation sequencing (NGS). SNP

array is conducted by using SNP marker probes that are designed for targeting specific genomic loci. For each SNP locus, two different probes are designed to target two possible types of alleles, while the signal intensities are measured by the total hybridization intensities for these two probes. More recently, massive parallel next-generation sequencing provides an appealing platform to sequence the whole genome at base-pair resolution. Applications of NGS usually include whole-exome sequencing (WES) and whole-genome sequencing (WGS). WES targets on scanning the protein-coding regions, which constitute 1% of the genome but attribute to about 85% of disease-related mutations (Petersen *et al.*, 2017). In contrast, WGS focuses on the entire genome, providing additional information in regulatory regions. Due to the lower cost and direct functional interpretation, WES is still a common practice for the identification of CNVs. Next, we will review the existing methods and tools for these two platforms.

1.2.1 CNV DETECTION WITH SNP ARRAY

Traditional SNP array genotyping technology enables the detection of CNVs with high resolution. Array-based CNV calling methods rely on modelling and examining data signals, including total signal intensity (referred to as log R Ratio [LRR]) and allelic intensity ratio (referred to as B allele frequency [BAF]) at each SNP. Specifically, for each SNP marker with two alleles referred to as A and B alleles, the raw intensities for A and B alleles are subject to normalization and generate normalized intensity X and Y (<https://icom.illumina.com/icom>). After calculating the total intensity $R = X + Y$ and relative intensity $\theta = \arctan(Y/X)/(\pi/2)$, LRR value is defined as $LRR = \log_2(R_{observed}/R_{expected})$, where $R_{expected}$ is computed by using linear interpolation

(Peiffer *et al.*, 2006). On the other hand, BAF that represents relative signal intensity ratio of the B allele can also be computed, referring to (Wang *et al.*, 2007).

After normalization of the derived signal intensities (i.e., LRR and BAF), there are mainly two types of statistical segmentation algorithms employed to perform CNV identification including change-point detection models (Truong *et al.*, 2018) and Hidden Markov models (HMM) (Seiser and Innocenti, 2014). The normalization is not the specific focus of our project, we will focus on reviewing the existing segmentation algorithms as below (Olshen *et al.*, 2004; Venkatraman and Olshen, 2007; Colella *et al.*, 2007; Wang *et al.*, 2007; Tibshirani and Wang, 2008; Niu and Zhang, 2012; Xiao *et al.*, 2015, 2019).

Change-point-based methods are a traditional statistical approach that has wide application in many areas, such as economics, neurology, manufacturing, and network (Lai, 1995; Lavielle and Teyssière, 2006; Koepcke *et al.*, 2016; Taylor and Letham, 2018; Kurt *et al.*, 2018). In the application to CNV identification, they perform hypothesis tests to exhaustively identify all the change-points that can partition the genome into segments with the same copy number. Circular binary segmentation (CBS) formulates a two-sample maximal student's t-test to recursively search for all the change-points, where the t-statistics compares the segment mean of LRR within a specific region to the mean of the remaining observations (Olshen *et al.*, 2004; Venkatraman and Olshen, 2007). A change-point is declared to be significant if the corresponding P-value is below a certain threshold (e.g., 0.05). The main drawback of CBS is its large computational burden. To solve this issue, Venkatraman and Olshen then proposed two strategies to improve the speed of the original CBS, including approximating the tail

probability of the maximal t-statistics and deriving an early stopping rule (Venkatraman and Olshen, 2007). More recently, Niu and Zhang proposed a novel local search segmentation procedure called the Screening and Ranking algorithm (SaRa), which is more computationally efficient than CBS (Niu and Zhang, 2012). The implementation of SaRa includes three steps, including calculation of mean difference before and after each point within a local window, finding the maximizer of the mean difference, and determining its statistical significance of being a change-point. This algorithm was utilized in the modSaRa method (Xiao *et al.*, 2017), which achieved higher detection power than CBS. However, the relatively allelic intensity information (i.e., BAF) is underutilized in modSaRa. Therefore, modSaRa2 (Xiao *et al.*, 2019) was further developed to integrate both LRR and BAF, which greatly improved the sensitivity and specificity of the method (Xiao *et al.*, 2019). Besides CBS and SaRa, another popular approach is the penalization method. Huang et al. formalized the change-point detection problem as a lasso variable selection problem, where each non-zero coefficient represents one breakpoint (Huang *et al.*, 2005). This framework also incorporates the genomic location information and spatial dependence through the penalty function. Tibshirani and Wang generalized the fused lasso framework to the change-point detection problem (Tibshirani and Wang, 2008). Compared to Huang’s method, it adds one more penalty term to ensure the sparsity of the change-point estimates.

Unlike change-point test-based methods, HMM-based methods consider a different model that the copy number status for all locations constitute an unknown sequence governed by Markov property. In this Markov model, the hidden state at each SNP locus only depends on the state of the previous SNP, where the probability of

observing a state change between nearby SNPs is described by a transition probability matrix. In addition to the transition probability matrix, HMM also constructs emission probability distributions to model both LRR and BAF, stating as the conditional distributions of data intensities given hidden states. HMM can then determine the most probable underlying sequence of the copy number states in a dynamic programming manner (Colella *et al.*, 2007; Wang *et al.*, 2007). Many HMM methods have been proposed in the past decades. QuantiSNP (Colella *et al.*, 2007) uses a series of copy number state-specific and distance-based functions to define transition probabilities in the HMM. In QuantiSNP, the emission probability matrix is formed as a mixture of normal and uniform distribution, and the uniform component is helpful in capturing data outliers. QuantiSNP also generates Bayes factors for all identified CNVs to determine their statistical significance. PennCNV (Wang *et al.*, 2007) utilizes similar transition and emission probabilities as QuantiSNP, but incorporates more sources of available data, such as family information and population genotype frequency, which can potentially improve its performance. genoCNV also develops the HMM in a similar way as PennCNV and QuantiSNP. The main novelty is that genoCNV not only identifies CNVs but also estimates their corresponding genotypes (Sun *et al.*, 2009). As discussed later, the HMM has also been widely used in CNV detection with WES.

1.2.2 CNV DETECTION WITH WES

WES CNV detection methods mainly use the read count data to identify CNVs. Specifically, for each targeted genomic region (e.g., exons), the read count value is generated by counting the number of reads that are mapped to this region. In general, WES CNV detection methods follow a two-step procedure: normalization and

segmentation. Due to the pervasive biases that are introduced during the sequencing experiment, normalization procedures are crucial in eliminating biases and restoring the true CNV signals. Following the data normalization step, statistical segmentation algorithms are applied to identify the segments that share the same underlying copy number. Most of the tools still use array-based detection algorithms, such as CBS and HMM described in Section 1.2.1. Many statistical methods have been developed and reviewed in literature, and we are discussing a few commonly used ones (Sathirapongsasuti *et al.*, 2011; Koboldt *et al.*, 2012; Krumm, Peter H. Sudmant, *et al.*, 2012; Fromer *et al.*, 2012; Plagnol *et al.*, 2012; Magi *et al.*, 2013; Jiang *et al.*, 2015; D’Aurizio *et al.*, 2016; Packer *et al.*, 2016; Jiang *et al.*, 2018).

ExomeCNV (Sathirapongsasuti *et al.*, 2011) is the earliest method for WES, which requires WES data from case-control pairs, such as matched tumor-normal samples. It calculates the read count ratios between cases and controls to reduce biases, assuming that biases are exon-specific and nearly constant across samples for a particular exon. CBS is then implemented to infer CNV calls in the segmentation. VarScan2 (Koboldt *et al.*, 2012) also needs WES data from case-control pairs. Unlike other methods, it does not use exon regions to generate read count data. Instead, VarScan2 constructs contiguous regions by binning the consecutive bases whose base-level read count ratios do not change significantly. Read count ratios are then calculated for those newly formed regions and further segmented by CBS. However, these methods all utilize control samples to remove biases, which cannot adjust technical variability between samples. Thus, to control sample-to-sample technical variability, ExomeDepth (Plagnol *et al.*, 2012) builds up a reference set that optimally chooses the number of samples to be added

to the reference set. After the construction of the optimal reference set, read count ratio data is calculated to adjust biases. ExomeDepth then uses HMM to generate CNV calls. However, all the above methods assumed that biases are only exon-specific. It was soon discovered that those biases are also sample-specific, which cannot be completely removed by using the case-control pairs. Therefore, CLAMMS (Packer *et al.*, 2016) selects a reference panel based on seven sequencing quality control metrics for each testing sample. Alternatively, CoNIFER (Krumm, Peter H Sudmant, *et al.*, 2012) andXHMM (Fromer *et al.*, 2012) uses SVD to extract the true CNV signals from the noisy data by removing the strongest latent factors that explain most of the data variations. The calculated SVD residuals are then analyzed by HMM to generate CNV calls. However, SVD-based approach cannot capture non-linear biases, such as GC content. More recently, CODEX (Jiang *et al.*, 2015) proposed a Poisson latent factor (PLF) model, which improves upon the SVD-based approach to remove both systematic and observed biases (i.e., GC content, exon length). Expected “null” coverages assuming there is no CNV are estimated and compared to the observed coverage. CODEX then uses CBS for the data segmentation. However, CODEX and SVD-based methods all lack sensitivity for detecting common CNVs, as the common signals will also be removed when removing top latent factors. Subsequently, CODEX2 improves CODEX, which utilizes CNV incident rate to rescue common signals (Jiang *et al.*, 2018). Nevertheless, both SVD and PLF methods process all samples at once, thus becoming computationally expensive with a large number of samples. EXCAVATOR (Magi *et al.*, 2013) and EXCAVATOR2 (D’Aurizio *et al.*, 2016) employ a three-step median normalization procedure to remove the effects from the GC content, mappability score and exon size. After bias correction,

logarithm ratio of normalized values between case and control sample are calculated. A shifting level model (SLM) algorithm is applied to perform segmentation, where SLM is similar to the HMM but with a different parametrization strategy (Magi *et al.*, 2010). The main difference between EXCAVATOR and EXCAVATOR2 is that EXCAVATOR2 also uses off-target read counts from WES data, which enlarges the detection spectrum.

1.2.3 GAPS IN THE CURRENT LITERATURE

Despite the progress in developing statistical methods, current methodologies for CNV detection still face analytical challenges. First, existing change-point detection-based methods are all developed with a strong independent assumption, ignoring the non-independence structure among genomic position in the genetics perspective. HMM-based segmentation methods use the Markov model with a state and distance-based transition probability matrix to model the dependence structure. However, HMM only employs a first-order Markov model, limiting the possibility to incorporate broad range correlations.

Second, conventional methods that we reviewed above all adopt a single-sample scanning strategy by simply applying the CNV calling algorithm to each sample repeatedly, which may result in high false discovery rates and low sensitivity due to sample-specific noise and CNV complexity. To address this challenge, several multiple sample-based methods have been developed that can utilize shared information across samples. Most of these methods still use single sample-based scan statistics (e.g., CBS statistics). After computing the single-sample statistics for all samples, multiple sample scan statistics is obtained by combining them together (N. R. Zhang *et al.*, 2010; Li and Tseng, 2011; Song *et al.*, 2016). Besides, Vert *et al.* implemented a group-lasso approach

to detect shared CNVs (Vert and Bleakley, 2010). However, this approach tends to be computationally expensive when applies to large-scale genetics data.

Third, CNV data genotyped by a single platform usually does not provide comprehensive coverage of the whole genome and is subject to platform-specific biases and nosiness. For example, WES merely explores the protein-coding regions, while SNP array can only genotype up to one million genetic variants. With the increasing availability of CNV data generated from different platforms (e.g., WES and SNP array). It is highly demanded to develop novel integrative CNV detection methods that can exploit all available information to boost both detection power and resolution. Zhou et al. (Zhou *et al.*, 2018) developed iCNV, which integrates SNP and WES data. iCNV presents significant detection power gain compared to single platform detection. However, it does not incorporate information across samples in its modelling framework, which tends to suffer from high false positive rate.

Thus, to address the limitations caused by the independence assumption, we proposed a novel algorithm (LDcnv) that models the correlated SNP data intensities (Chapter 2). Meanwhile, we also proposed a correlation-based method, CORRseq, as a novel release of LDcnv in analyzing correlated WES data (Chapter 3). Moreover, to address the obstacle due to the sparsity nature of SNP array and WES data and improve on both detection accuracy and resolution, we developed BMI-CNV, a Bayesian Multi-sample and Integrative CNV calling method, which can efficiently integrate data from multiple platforms (i.e., WES and SNP array) and multiple samples (Chapter 4).

CHAPTER 2

INTEGRATING GENOMIC CORRELATION STRUCTURE IMPROVES COPY NUMBER VARIANTS DETECTION¹

2.1 ABSTRACT

Existing CNV detection algorithms have been developed with a strong assumption of independent observations in the genetic loci, and they assume each locus has an equal chance to be a change-point. However, this assumption is violated in the genetics perspective due to the existence of correlation among genomic positions such as linkage disequilibrium (LD). In this study, we provided theoretical proof to demonstrate the correlation structure of CNV data generated from SNP array. Motivated by this evidence, we developed a novel SNP array CNV calling algorithm, LDcnv, that models the CNV data with its biological characteristics relating to the genomic dependence structure. To evaluate the performance of LDcnv, we conducted extensive simulations and analyzed large-scale HapMap datasets. We showed that LDcnv presented higher precision in detecting short CNVs compared to existing methods.

2.2 INTRODUCTION

Technically, the detection of CNVs is the finding of breakpoints or boundaries of copy number regions from the genotyping signals. Change-point tests have been commonly used and implemented in several software and tools for chromosomal segmentation

¹ Xizhi Luo, Fei Qin, Guoshuai Cai and Feifei Xiao. *Bioinformatics*. 2021 Apr 20;37(3):312-317. Reprinted here with permission of the publisher.

(Darvishi, 2010; Deng, 2011; Gai, et al., 2010). Among them, circular binary segmentation (CBS) has been widely used and is based on an exhaustive test (Olshen, et al., 2004). More recently, a novel segmentation procedure was utilized in modSaRa that adopted a local search strategy and was efficient for whole genome analysis with low computational complexity (Niu and Zhang, 2012; Xiao, et al., 2019; Xiao, et al., 2017). Our team then further proposed modSaRa2 that integrates Log R Ratio (LRR) and B allele Frequency (BAF) that greatly improved the sensitivity and specificity (Xiao, et al., 2019). Nevertheless, all of these algorithms were developed with a strong assumption of independent observations across the genetic loci and they assume each locus has an equal chance to be a breakpoint (i.e., boundary of CNVs). However, this assumption is violated in the genetics perspective given the non-independence structure among genomic positions, which is referred to as linkage disequilibrium (LD). Dictated by the presence of recombination hotspots that segment the genome into separate blocks, LD describes the non-independent transmission of the alleles at adjacent locations in the genome. Moreover, it was discovered early that CNVs are outcomes of evolution and they originated from recombination-based processes (Zhang *et al.*, 2009), which demonstrated the possibility of the existence of CNV breakpoints located at the recombination hotspots, which violates the assumption of previous segmentation methods as mentioned above. Given this evidence, it is essential to integrate the biological characteristics into statistical modeling for CNV detection.

Motivated by this fact, we here develop an accurate and fast segmentation algorithm, referred to as LDcnv, by modeling the genomic correlation structure with a local search strategy for optimized computational efficiency, with the correlation

structure derived as a function of LD measures. To investigate the performance of the newly proposed algorithm, we conducted simulation studies to investigate its performance in SNP array studies in a variety of scenarios. We demonstrated the improved performance of the method in array-based real data analysis by using a set of “gold standard validation sets” from the HapMap projects (McCarroll *et al.*, 2008; Conrad *et al.*, 2010; D. Altshuler *et al.*, 2010). Overall, the new algorithm presented high sensitivity and accuracy in CNV detection, especially for detection of small-sized CNVs.

2.3 METHODS

2.3.1 THEORETICAL DERIVATION: CORRELATION STRUCTURE IN CNV DATA

In this study, we hypothesized that the integration of the genomic correlation structure will boost the accuracy of CNV detection. In this section, we therefore provide theoretical evidence to support that the genetic intensity data (i.e., LRR) are non-independent and presenting a correlation structure that should be deliberated in statistical modeling for CNV detection.

We start from the generation of the SNP array intensities LRR, which has been introduced in Section 1.2.1. In our study, to present the correlation of two adjacent bi-allelic SNPs, we assume the reference allele and alternative allele were A and a for the first SNP, and B and b alleles for the second SNP, alleles A and B occur with population frequency p_A and p_B . The total signal intensities for the two alleles are therefore $X_A + Y_A$ and $X_B + Y_B$. Under the Hardy-Weinberg equilibrium assumption (HWE), the joint probability for the nine possible genotypes can be calculated (shown in Supplementary Table A.1). For example, for the genotype $AABB$, the genotype frequency will be

$(p_A p_B + D_{AB})^2$ where D_{AB} is the coefficient of linkage disequilibrium between the two SNPs.

Our main interest is then to calculate the correlation of $LRR = \log_2(R_{observed}/R_{expected})$ between the two SNPs. After applying the Taylor expansions, the correlation of the LRR intensities can be approximately represented by the correlation of $R_{observed,A}$ and $R_{observed,B}$, which is expressed by

$$\rho_{AB} = \frac{cov(X_A + Y_A, X_B + Y_B)}{\sqrt{var(X_A + Y_A)var(X_B + Y_B)}} \quad (1)$$

Derivation of the ρ_{AB} shows the existence of correlation structure of the LRR intensities across the genome, which will be further discussed in the results (Section 2.4.2).

2.3.2 LDcnv ALGORITHM

Given the evidence of correlation structure in the dataset, we further proposed a novel algorithm constructed from a basic change-point model which identifies underlying mean shifts in intensities and locate breakpoints along the genome. Specifically, let $\mathbf{Y} = (Y_1, \dots, Y_M)^T$ denote the genetic intensities for a genome sequence with M biomarkers (i.e., SNPs in array data or exons in WES data). A high-dimensional linear normal mean model is usually adopted to model the signal intensities in previous literature and is also used in our framework (Niu and Zhang, 2012; Xiao *et al.*, 2017, 2019),

$$Y_i = \mu_i + \varepsilon_i, i = 1, 2, \dots, M. \quad (2)$$

where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_M)^T$ is the underlying piecewise constant mean vector, $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_M)^T \sim N_p(0, \boldsymbol{\Sigma})$ are the error terms. The breakpoints or boundaries of the CNVs are position τ 's that $\mu_\tau \neq \mu_{\tau+1}$, where the locations of these breakpoints are reflected in the model as change points. We assume that there are T ordered change points in the

sequence, $0 < \tau_1 < \dots < \tau_T < M$. Thus, the main goal is to estimate the location vector consisted of all the change points, $\boldsymbol{\tau} = (\tau_1, \dots, \tau_T)^T$. Studies have worked on the problem of identifying the location of breakpoints when the Y s are independent. In this chapter, to capture the biological characteristics (i.e., genetic correlation) in the process of copy number states inference, we proposed the algorithm, referred to LDcnv, that systematically integrated the genomic correlation structure of the genetic intensities into statistical modelling. We assume the genetic intensities follow a multivariate normal distribution given the dependence structure of the genome.

$$Y \sim MVN(\mu, \Sigma) \quad (3)$$

where Σ is the covariance matrix with dimension $m \times m$. The covariance matrix (Σ) can be estimated by using the correlation matrix estimated from the samples or an exterior large-scale population reference dataset such as samples from the 1000 Genomes project (Auton, Gonçalo R. Abecasis, *et al.*, 2015).

Similar to the SaRa algorithm (Niu and Zhang, 2012), LDcnv was also built with the local diagnostic function $D(x)$ defined as the average mean difference in the observations before and after a point x ,

$$D(x, w) = \sum_{k=1}^w Y_{x+1-k}/w - \sum_{k=1}^w Y_{x+k}/w \quad (4)$$

where w is the pre-defined bandwidth. In our algorithm, the quantity of $D(x, w)$ depends on the local $2w$ data points $\tilde{Y} : Y_{\tau+1-w}, \dots, Y_{\tau}, \dots, Y_{\tau+w}$ where $\tilde{Y} \sim MVN(\tilde{\mu}, \tilde{\Sigma})$. $\tilde{\mu}$ is a sub-vector of μ with length $2w$; $\tilde{\Sigma}$ is a sub diagonal matrix of the covariance matrix Σ with dimension $2w \times 2w$, respectively. Then $D(x)$ can be rewritten as $D(x) = \tilde{a}\tilde{Y}$. \tilde{a} is a $2w$ vector takes the form $\frac{1}{w}[\mathbb{1}_{w \times 1} \quad -\mathbb{1}_{w \times 1}]$. Instead of using an independence correlation

matrix as used in SaRa, our algorithm efficiently incorporates the local correlation information into the statistical modeling without sacrifice on computational efficiency.

By derivation with the linear property of multivariate normal distribution, we obtained $D(x) \sim N(\vec{a}\tilde{\mu}, \vec{a}\tilde{\Sigma}\vec{a}^T)$. It turned out that the distribution of the local diagnostic function became a univariate normal with a covariance matrix depending on the local information, $\tilde{\Sigma}$. Since both $\tilde{\mu}$ and $\tilde{\Sigma}$ are known or can be estimated from the local data sequence, the mean and variance of $D(x)$ are functions of bandwidth w . The bandwidths w should be carefully chosen, referring to (Xiao et al., 2015), and we suggest using multiple bandwidths in applications, which has been implemented in our LDcnv method and R package (<https://github.com/adamluo12/LDcnv>).

2.3.3 COPY NUMBER INFERENCE

After calculation of the local diagnostic statistic $D(x)$, hypothesis testing is implemented to find change-point candidates by a local screening and ranking strategy (Niu and Zhang, 2012). A similar strategy has been used in our previous work (Xiao et al., 2017, 2019), which guarantees high computational speed in a whole genome scan.

Providing the distribution of $D(x)$, we first define the w -local maximizer of a function. For any data point x , the interval $(x - w, x + w)$ is called the w -neighborhood of x . And x is a w -local maximizer of function $f(\cdot)$ if

$$f(x) \geq f(x') \text{ for all } x' \in (x - w, x + w) \quad (5)$$

Then let \mathcal{L} be the set of all local maximizers of the function $|D(x, w)|$ and we can select a subset $\hat{\mathcal{M}} = \{\hat{\tau}_1 < \hat{\tau}_2 < \dots < \hat{\tau}_{\hat{M}}\} \subset \mathcal{L}$ by setting a threshold $|D(\hat{\tau}, w)| > \gamma$, where $\hat{\mathcal{M}}$ and \hat{M} are the estimators for the locations and the number of change-points, respectively.

To set up the threshold γ , we adopted a multiple testing-based method proposed by Hao

et al., which focused on assessing and controlling the false discovery rate (FDR) for change-point location estimators (Hao *et al.*, 2013). This FDR approach was demonstrated to be effective in selecting thresholding parameter that controlled the FDR at a target rate (e.g., FDR=0.05). See Theorem 1 in (Hao *et al.*, 2013) and Supplementary B.1 for more details. As a result, the local maximizers $\hat{\mathcal{M}}$ or, equivalently, local minimizers of p-values were selected.

Then we used the modified Bayesian Information Criteria (mBIC) to further eliminate false positives as proposed in (Zhang and Siegmund, 2007):

$$mBIC(\tilde{\mathcal{M}}) = \frac{n}{2} \log(\hat{\sigma}_{\tilde{\mathcal{M}}}^2) + \tilde{J} \log(n) + \frac{1}{2} \sum_{i=1}^{\tilde{\mathcal{M}}+1} \log\left(\frac{x_{(i)}}{n} - \frac{x_{(i-1)}}{n}\right) \quad (6)$$

where $\tilde{\mathcal{M}}$ is all the possible values of $\hat{\mathcal{M}}$ and $\hat{\sigma}_{\tilde{\mathcal{M}}}^2$ is the maximum likelihood estimator of the variance assuming $x_1, \dots, x_{\tilde{\mathcal{M}}}$ are change points. Then the final estimated number and the locations of change points are $\hat{\mathcal{M}}' = \operatorname{argmin}(mBIC(\tilde{\mathcal{M}}))$ and $\hat{\mathcal{M}} = \{\hat{t}_1 < \hat{t}_2 < \dots < \hat{t}_{\hat{\mathcal{M}}'}\}$, respectively. For copy number inference, Gaussian mixture model-based clustering was used for copy number state classification (Xiao *et al.*, 2015). Each segmented region will be classified using a three-state classification scheme (deletion, normal/diploid and duplication).

2.3.4 NUMERICAL SIMULATION STUDIES

With the new proposed algorithm, we conducted extensive simulations to evaluate the performance in CNV identification. To simulate correlated genomic intensities, we used the first-order autoregressive (AR1) process:

$$Y_i = c + \phi Y_{i-1} + \varepsilon_i, i = 1, 2, \dots, n, \quad (7)$$

where Y_i was the intensities for the i -th marker; ε_i was a Gaussian white noise process with mean zero and variance σ_ε^2 ; ϕ was a known coefficient that controlled the autocorrelation of the data series (for example, $|\phi| < 1$ generates a stationary sequence); c was a constant and n was the total number of markers. The underlying mean μ , variance $var(Y_i)$ and auto-covariance B_n were given as: $\mu = \frac{c}{1-\phi}$, $var(Y_i) = \frac{\sigma_\varepsilon^2}{1-\phi^2}$ and $B_n = \frac{\sigma_\varepsilon^2}{1-\phi^2} \phi^{|n|}$. Using the AR1 process allows to flexibly adjust the distribution of the data by changing the underlying distribution of the white noise term and generates data much faster than the multivariate normal distribution assumption-based process without decomposing the covariance matrix.

We randomly generated LRR and BAF intensities for 100 sequences (i.e., chromosomes) with 20,000 markers. For each sequence, 40 dispersed and non-overlapping CNV segments were generated, the locations of which were randomly selected. The mean and variance were empirical values provided by the Illumina website (<https://www.illumina.com/documents>). We constructed different scenarios with different combinations of CNV sizes, states and correlation levels. The CNV sizes varied from 10~50 markers, 50~100 markers and 100~200 markers. The CNV status included deletion of a single copy (Del.s), deletion of double copies (Del.d), duplication of a single copy (Dup.s) and duplication of double copies (Dup.d). To investigate the CNV data with different levels of correlations, the value of ϕ which was equivalent to the Pearson's correlation coefficient in theory, was set to be 0.1, 0.3 and 0.5, corresponding to weak, moderate and strong correlation strength, respectively. With the generated CNV data, we compared the proposed method to an independence assumption-based method, CBS, a

shifting level model based method, SLMSuite and a hidden Markov model based method, PennCNV (Olshen *et al.*, 2004; Wang *et al.*, 2007; Orlandini *et al.*, 2017). For LDcnv, we also compared the performance of using LRR data only (LDcnv_LRR) and using the estimates that integrates LRR and BAF (LDcnv) (Supplementary B.5) (Xiao *et al.*, 2019). The performance of these methods was demonstrated by comparing the true positive rate (TPR) and false positive rate (FPR).

2.3.5 PERFORMANCE EVALUATION BY APPLICATION TO THE HAPMAP DATASETS

To evaluate the proposed LDcnv algorithm, we analyzed 180 healthy individuals, which are all Utah residents with Northern and Western European ancestry (CEU). The SNP array data were downloaded from the international HapMap 3 Consortium (D. Altshuler *et al.*, 2010). The genetic intensities were then pre-processed by the `genomic_wave.pl` function of PennCNV (Wang, et al., 2007) to adjust the GC wave content effects. To further assess the performance of LDcnv, we used the CNV profiles that have been validated experimentally or statistically in three previous microarray studies (McCarroll *et al.*, 2008; Conrad *et al.*, 2010; D. Altshuler *et al.*, 2010). The HapMap project utilized stringent genotyping quality control (QC) and merged results from multiple calling algorithms, which produced 856 high-quality CNV calls (D. Altshuler *et al.*, 2010). McCarroll et al. identified 1,320 high resolution CNV calls by joint analysis of multi-platforms data (McCarroll *et al.*, 2008), whereas Conrad et al. used tiling oligonucleotide array to generate a map of 11,700 CNVs, among which 8,599 were independently validated through stringent validation procedures (Conrad *et al.*, 2010). Specifically, stringent QC procedures were adopted (i.e., the reported CNVs must overlap with at least

2 SNP markers, have less than 5% missing rate across all samples) to construct the “gold standard validation sets”.

Using this “gold standard validation sets”, we compared the performance of the LDcnv method against PennCNV, SLMSuite and CBS (Olshen *et al.*, 2004; Wang *et al.*, 2007; Orlandini *et al.*, 2017). For CNV calls from all four methods, to obtain high-quality CNV profiles, we excluded CNVs with less than ten markers in the calling results. Besides, we used the database of genomic variants (DGV) as a reference of common variants to select the high-quality CNV profile, which curates CNV records from 55 independent studies of clinically normal populations (MacDonald *et al.*, 2014).

These methods were assessed by the precision rate, recall rate and F1 score measures. The precision rate was defined as the ratio of identified true positives over the total number of identified CNVs. The recall rate was the ratio of identified true positives over the total number of “true CNVs” in the “gold standard validation sets”. The F1 score was defined as the harmonic mean of precision and recall rate which reflected the overall accuracy. Moreover, we evaluated the performance in subsets of the validation sets which consist of CNV records spanning less than 10 markers to assess the performance of these methods in detecting short CNVs.

2.4 RESULTS

2.4.1 REAL DATA SHOWS THAT THAT CNV LOCATIONS ARE RELATED TO THE GENOMIC STRUCTURE

To explore the relationship between CNV locations and LD structure, we utilized the real data high-quality CNV profile from the international HapMap 3 Consortium which merged probe-intensity data from both Affymetrix and Illumina arrays (D. Altshuler *et*

al., 2010). The profile set contained 856 CNV records from 1,184 individuals. We randomly selected 300 high-quality CNVs and mapped to the LD block maps (Supplementary Figure A.1). Obviously, most of these CNV breakpoints are located outside of the LD blocks (across block, hybrid or random), with only 2.0% residing within LD blocks (inter-block). Among the CNVs that involve LD structure (i.e., across block, inter-block and hybrid), only 11.0% were spanning at least one LD blocks. These results implicated that CNVs are not randomly distributed across the genome, and their distribution of the breakpoints is closely related to the local LD structure. Such results motivated the development of our LDcnv algorithm and are consistent with the theoretical derivation of the correlation structure of genetic intensities in SNP array data (Section 2.4.2).

2.4.2 THEORETICAL DERIVATION REVEALED THE CORRELATION STRUCTURE IN GENETIC INTENSITIES

First, to further demonstrate the necessity of integrating the correlation structure into the segmentation algorithm, we initiated the study by a theoretical derivation to demonstrate the correlation structure of the genetic intensities (i.e., LRR) between two adjacent genomic loci. As a continued discussion of Section 2.3.1, we have the correlation coefficient of LRR between two loci expressed as $\rho_{AB} = \frac{cov(X_A+Y_A, X_B+Y_B)}{\sqrt{var(X_A+Y_A)var(X_B+Y_B)}}$, in which X and Y are the normalized signal intensities of the two alleles in a SNP (e.g., A and a).

For $cov(X_A + Y_A, X_B + Y_B)$, we obtained,

$$cov(X_A + Y_A, X_B + Y_B) = cov(X_A, X_B) + cov(X_A, Y_B) + cov(Y_A, X_B) + cov(Y_A, Y_B) \quad (8)$$

As $cov(X_A, X_B) = E(X_A X_B) - E(X_A)E(X_B)$, the expected values of the normalized signal intensities X_A, X_B and the expected values of their product need to be derived. We assume the joint probability density function to be $f_{X_A, X_B}(x_A, x_B)$ which are bivariate normal distributions conditional on the genotype G :

$$f_{X_A, X_B}(x_A, x_B) = \sum_{k=1}^4 f_{X_A, X_B}(x_A, x_B | G) P(G = G_k) \quad (9)$$

where $G = [AABB, AABb, AaBB, AaBb]^T$ is the vector of genotypes that contain alleles A and B . After mathematical derivation (detailed in Appendix B.2), the covariance between the two normalized signal intensities can be formulated as:

$$cov(X_A, X_B) = \sum_{k=1}^4 E(X_A | G_k) E(X_B | G_k) [P(G = G_k) - q(G = G_k)]. \quad (10)$$

$q(G = G_k)$ is the genotype frequency under the condition that the two loci are in LD. For example, $q(AABB) = p_A^2 p_B^2$. The expression of all the other genotype frequencies $P(G = G_k)$ can be found in Supplementary Table A.1.

Similarly, we can derive the other three terms in equation (8) and then obtain $cov(X_A + Y_A, X_B + Y_B)$ as

$$\sum_{i=1}^4 \sum_{k=1}^4 E(X_A | G_{ik})^{I_{1i}} E(X_B | G_{ik})^{I_{2i}} E(Y_A | G_{ik})^{1-I_{1i}} E(Y_B | G_{ik})^{1-I_{2i}} [P(G_{ik}) - q(G_{ik})] \quad (11)$$

where $G_1 = [AABB, AABb, AaBB, AaBb]^T$, $G_2 = [AABb, AAbb, AaBb, Aabb]^T$, $G_3 = [AaBB, AaBb, aaBB, aaBb]^T$ and $G_4 = [AaBb, Aabb, aaBb, aabb]^T$. I_{1i} and I_{2i} are indicator functions of whether X_A and X_B contribute to the bivariate density in equation (9). $I_{1i} = 1$, if $i = 1$ or 2 . $I_{2i} = 1$, if $i = 1$ or 3 . Otherwise, $I_{1i} = I_{2i} = 0$.

Combining results from equation (11) and the expression of the denominator of ρ_{AB} in equation (7) from Appendix B.3, the correlation of LRR between the two loci can be defined as:

$$\rho_{AB} = \frac{\sum_{i=1}^4 \sum_{k=1}^4 E(X_A|G_{ik})^{I_{1i}} E(X_B|G_{ik})^{I_{2i}} E(Y_A|G_{ik})^{1-I_{1i}} E(Y_B|G_{ik})^{1-I_{2i}} [P(G_{ik}) - q(G_{ik})]}{\sqrt{\pi_1 \text{var}(X_A) + \pi_2 \text{var}(Y_A)} \sqrt{\pi_3 \text{var}(X_B) + \pi_4 \text{var}(Y_B)}}$$

where G_{ik} denotes the underlying k -th genotype contained in the i -th bivariate normal distribution (details in Appendix B.2); $P(G_{ik})$ is the corresponding genotype frequencies under the HWE assumption; $q(G_{ik})$ is the genotype frequencies assuming independent loci. According to expression of above Equation, the correlation of LRR depends on the association of the two SNPs which was measured by the LD coefficient D_{AB} . For example, $P(G_{ik}) - q(G_{ik}) = (p_A p_B + D_{AB})^2 - p_A^2 p_B^2$ for genotype $AABB$ ($i = 1, k = 1$). As we see, the correlation of the LRR will be equal to zero if the LD coefficient between the two SNPs equals to zero since $\rho_{AB} = 0$ if $D_{AB} = 0$.

As such, in theory, we showed that the correlation of LRR intensities is related to the coefficient of LD measure between two SNPs, although the relationship does not admit a simple format. This result demonstrated the correlation structure in SNP array data and imply the need to take the correlation structure of LRR into consideration in CNV detection.

2.4.3 SIMULATION STUDIES SHOW IMPROVED PERFORMANCE OF LDcnv

First, we used the simulated data to evaluate the performance of the LDcnv methods in SNP array analysis under a variety of scenarios: (1) different correlation levels (i.e. $\phi = 0.1, 0.3$ and 0.5); (2) different CNV sizes (i.e. 10~50 markers, 50~100 markers and 100~200 markers); and (3) different CNV states (i.e. Del.d, Del.s, Dup.d and Dup.s). The

LDcnv method presented a consistent power across various settings. For data with moderate correlation coefficient ($\phi = 0.3$) that is assumed to be close to real data, we found gain in detecting CNVs from single copy duplication/deletions (i.e., Dup.s and Del.s) to double copy changes (i.e., Dup.d and Del.d) (Table 2.1), and the estimations of CNVs are stable across different scenarios of CNV states and sizes. For short CNVs (<50 markers), the performance of the LDcnv was obviously superior to the other methods when the CNVs had small jump sizes (Dup.s and Del.s), whereas the CBS and SLMSuite methods showed diminished power. For example, when the CNVs had a length between 10-50 markers and the CNV state was single copy duplication (Dup.s), the LDcnv method had a TPR at 0.97, while the FPR was 0.03. The corresponding TPRs and FPRs for PennCNV were 0.91 and 0.07; 0.85, 0.11 for CBS; and 0.77, 0.04 for SLMSuite, respectively. When the CNV size increased from 10-50 markers to 100-200 markers, the LDcnv maintained a stable performance and was comparable to the SLMSuite method, except for detecting single copy duplications. In the contrary, the PennCNV method presented diminished power. For example, for single copy deletions with a length between 100 to 200 markers, the LDcnv method (TPR=0.99, FPR=0.03) and SLMSuite (TPR=0.99, FPR=0.01) showed significantly better performance than PennCNV (TPR=0.86, FPR=0.11). A similar pattern was observed when the correlation increased from 0.1 to 0.5 (Supplementary Tables A.3 and A.4). Besides, increased sensitivity and specificity of LDcnv (that integrates LRR and BAF) were clearly observed compared to LDcnv that only uses LRR intensities (LDcnv_LRR), which was consistent with our previous findings (Xiao *et al.*, 2019).

In conclusion, the LDcnv method, which integrated the correlation structure in the model, largely presented overall high accuracy, stability, and robustness in CNV detection, especially for detection of short CNVs with small jump sizes.

2.4.4 APPLICATION TO THE HAPMAP DATASETS

We further applied the LDcnv method to a real data study in comparison with CBS, SLMSuite and PennCNV (Olshen *et al.*, 2004; Wang *et al.*, 2007; Orlandini *et al.*, 2017). Using the DGV as a common variant reference database, 47.30% of the CNVs identified by LDcnv have been reported as common variants that are not diseases relevant.

With the validation sets from the three datasets (including HapMap3, Conrad *et al.*, and McCarroll *et al.*), the total number of “true” CNVs included in these three datasets were 19,936, 121,453 and 11,961, separately. Among them, 10,005 (50.18%), 98,387 (81.01%) and 5,277 (44.12%) were short CNVs with length less than ten markers. The overall performance of the LDcnv method was greater than that of the other methods in all three validation sets (Supplementary A.5, Figure 2.1). Specifically, LDcnv presented the highest F1 scores and detected much more true positives than other methods, although the precision is compromised. For example, in the HapMap dataset, LDcnv accurately detected 7,016 true positives. The corresponding true positives were 3,760 for PennCNV; 3,965 for CBS and 4,942 for SLMSuite. For detection of short CNVs (Supplementary Table A.6, Supplementary Figure A.2), the LDcnv method was also superior to the other methods in all three validation sets and it presented obviously the highest F1 scores and identified the largest number of true positives in detecting short CNVs. As expected, PennCNV was the conservative one that detected the lowest number therefore presented the lowest precision rate (Supplementary Table A.6, Supplementary Figure A.2). These

results further demonstrated that the integration of correlation structure significantly improved the overall performance of CNV detection in LDcnv.

2.5 DISCUSSION

In this chapter, we first presented the theoretical derivation of the correlation structure of the genetic intensity data from SNP array data. We found that the array-based LD structure, which was computed from the genotype frequencies, can be reflected in the correlation structure of the genetic intensity data. We stated that the correlation between two loci in the genetic intensity will depend on the LD coefficient computed from SNP allele frequencies. This evidence provided strong support of the existence of genomic dependence structure in the CNV data. A correlation-based segmentation algorithm for CNV detection that accommodates the non-independence nature of the genetic intensities is then introduced. Simulation and real data analyses suggested that the LDcnv algorithm presented stable performance and essential advantages over the other comparative methods, especially the independence assumption based methods.

The largest power gain tended to occur when CNVs were short and with small jump sizes, e.g., the duplication of a single copy. The superiority of the LDcnv algorithm over PennCNV and SLMSuite was further demonstrated, especially in detecting short CNVs, which is the most difficult copy number states to be detected due to the embedded undetectable signal in the random noises. A possible explanation for this phenomenon is that short CNVs tend to have more evident correlation structure when they are located within an LD block. Such a characteristic cannot be easily captured by the hidden Markov model adopted in PennCNV and SLMSuite, which assumes a constant level of Markov dependence across the genome.

Indeed, the clinical relevance of small CNVs has been demonstrated in many studies. For example, Reza et al. (Asadollahi *et al.*, 2014) investigated a cohort of 714 patients with neurodevelopment disorders and verified the diagnostic importance of small CNVs. However, due to the noise of genotyping data, small segments are usually very difficult to distinguish from the normal noise signals. As such, the LDcnv algorithm may serve as an important tool for detecting small CNVs.

With LDcnv, we address the non-independence noise signal assumption by introducing a covariance matrix in the statistical modeling. To be noted, although this proposed algorithm was motivated by the existence of LD structure in the genome, our statistical modeling was not utilizing such information directly since the LD coefficient was computed from the genotypes instead of from the genetic intensities. To retain the covariance structure of the genetic intensities in the model, we can either use the correlation matrix estimated from the samples in the data or LD-based computation from reference samples (i.e., samples from the 1000 Genomes project). The advantage of the data-based estimation of the covariance relies on its feasibility and simplicity; however, the covariance might be data specific, and the computational concerns will be encountered for large sample sizes. In contrast, the LD-based estimates with information coming from the population level might be more stable but be susceptible to a specific population substructure. As discussed in Mathew et al. (Mathew *et al.*, 2018), an alternative way is to use the map functions (e.g., the Haldane function) in an exponential function to estimate the covariance structure on each chromosome.

Moreover, we have mainly demonstrated that the LD structure can be reflected in the genetic intensity data (i.e., LRR in SNP array). However, the correlation structures of

the BAF intensities were still not clear and not easily constructed. As a result, the theoretical study of the method that integrate LRR and BAF requires further studies. To alleviate the problems brought by such limitation, we only adopted data-based estimates for the covariance structure of the integrated LRR and BAF intensity in the statistical modeling. In addition, we used 300 high-quality CNV data from HapMap to demonstrate that CNVs were not randomly distributed across the genome. One possible explanation might be that the LD blocks under study were not long enough to contain a CNV. Rigorous examinations of this assumption with shorter CNVs should be further studied in the future.

To profile CNV, various techniques have been used including SNP array and next generation sequencing (NGS) technologies. Due to the high cost and prohibitive computational requirements in NGS, SNP array is still an excellent choice for a genome-wide analysis of CNVs in large GWASs, and there are many unexplored large-scale cohorts with SNP array data. This work is motivated to fill in a gap in the statistical modeling of CNV data and we start with SNP array data analysis as one of the important directions of applications. Our method developed in this study also has great potential to be implemented in the WES data analysis, which we pursue in chapter 3.

TABLES AND FIGURES

Table 2.1 Summary of CNV calls on simulated data at $\phi = 0.3$ from all methods. True positive rates (TPRs) and false positive rates (FPRs) of LDcnv, LDcnv_LRR, PennCNV, SLMSuite and CBS with different CNV states and CNV sizes, the autoregressive coefficient (ϕ) was fixed at $\phi = 0.3$ which was corresponding to Pearson's correlation coefficient at 0.3. Del.d: deletion of double copies; Del.s: deletion of single copy; Dup.s: duplication of single copy; Dup.d: duplication of double copies.

CNV State	Method	CNV length (markers)					
		10~50		50~100		100~200	
		TPR	FPR	TPR	FPR	TPR	FPR
Del.d	LDcnv	0.99	<0.01	0.97	<0.01	0.99	0.01
	LDcnv_LRR	0.98	0.01	0.98	0.01	0.98	0.01
	PennCNV	1.00	<0.01	1.00	<0.01	1.00	<0.01
	SLMSuite	0.99	<0.01	1.00	<0.01	0.99	<0.01
	CBS	1.00	0.04	1.00	0.07	1.00	0.09
Del.s	LDcnv	0.99	0.02	0.97	0.01	0.99	0.03
	LDcnv_LRR	0.98	0.01	0.98	0.01	0.98	0.02
	PennCNV	0.96	0.03	0.95	0.04	0.86	0.11
	SLMSuite	0.98	0.01	0.99	0.01	0.99	<0.01
	CBS	0.98	0.03	0.98	0.04	0.98	0.05
Dup.s	LDcnv	0.97	0.03	0.94	0.03	0.93	0.05
	LDcnv_LRR	0.92	0.08	0.92	0.09	0.92	0.12
	PennCNV	0.91	0.07	0.92	0.08	0.87	0.11
	SLMSuite	0.77	0.04	0.95	0.04	0.97	0.03
	CBS	0.85	0.11	0.88	0.12	0.88	0.13
Dup.d	LDcnv	1.00	<0.01	1.00	<0.01	0.99	<0.01
	LDcnv_LRR	1.00	<0.01	1.00	<0.01	1.00	<0.01
	PennCNV	1.00	<0.01	0.99	<0.01	0.99	0.01
	SLMSuite	1.00	<0.01	1.00	0.00	1.00	0.00
	CBS	1.00	0.01	1.00	0.02	1.00	0.04

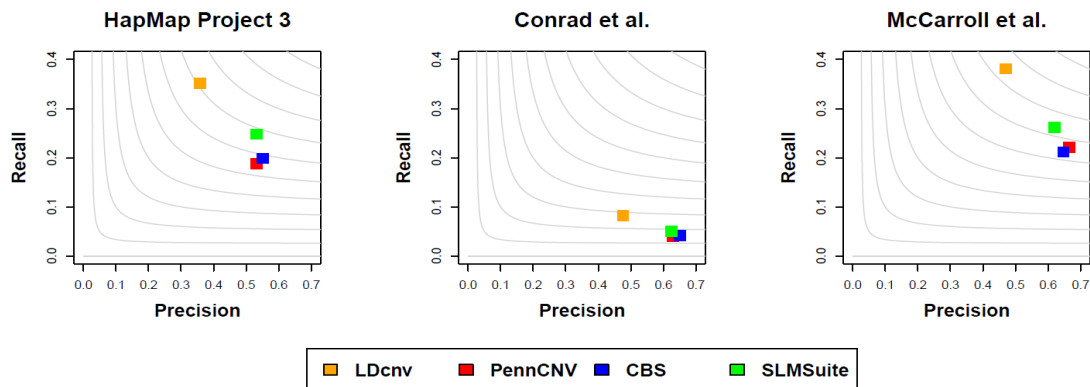


Figure 2.1 Assessment of CNV calls generated by LDcnv, PennCNV, CBS and SLMSuite methods with validation datasets from (a) HapMap 3 (b) Conrad et al (c) McCarroll et al. The x-axis is the precision and the y-axis present the recall rate. The grey contours are F1 scores calculated as the harmonic mean of precision rate and recall rate, which are 0 to 1 from left bottom to right top.

CHAPTER 3

INTEGRATING GENOMIC CORRELATION STRUCTURE ENHANCES COPY NUMBER VARIANTS DETECTION FROM WHOLE-EXOME SEQUENCING DATA²

3.1 ABSTRACT

WES is effective in identifying CNVs with finer resolution compared to SNP array. We have previously shown evidence of the genomic correlation structure in SNP array data and developed a novel algorithm, LDcnv, which showed significantly improved detection power through integrating the correlation. However, it remains unexplored whether the genomic correlation exists in WES data and how such correlation structure integration can improve CNV detection accuracy.

As a continuation of Chapter 2, in this project, we first used the 1000 Genomes Project WES data to evaluate the correlation structure before and after two commonly used normalization approaches: a median-based method and a Poisson latent factor (PLF) model-based method. Strong evidence of correlation structure was found before and after implementing the median normalization. Motivated by this fact, we developed a correlation-based method, CORRseq, as a novel release of the LDcnv algorithm in profiling WES data. Furthermore, given the weak correlation structure of WES data after PLF-based normalization, we also proposed SARA for analyzing independent data.

² Fei Qin & Xizhi Luo, Guoshuai Cai and Feifei Xiao. *Briefings in Bioinformatics*, 2021; bbab215. Reprinted here with permission of the publisher.

The performance of CORRseq and SARaseq was evaluated in extensive simulation studies and real data from the 1000 Genomes Project. CORRseq outperformed existing methods in detecting medium and large CNVs. While SARaseq, which assumed the independent observations, performed the best in detecting short CNVs. The application of CORRseq to the TRICL consortium identified candidate lung cancer risk associated CNVs in 1p21.1, 1p23.3, 11q11.2, 17q21.32, 22q11.22 regions.

3.2 INTRODUCTION

With the rapid technological development, WES has offered an appealing platform for CNV identification due to its low cost, high coverage, and relatively simplified analysis process (Zare *et al.*, 2017). Many statistical methods have been developed to detect CNVs from WES data, such asXHMM (Fromer *et al.*, 2012), CoNIFER (Krumm, Peter H Sudmant, *et al.*, 2012), EXCAVATOR (Magi *et al.*, 2013), and CODEX (Jiang *et al.*, 2015) (see Section 1.2.2 for a detailed review). Technically, these methods consist of two main procedures: data normalization and segmentation.

WES data is highly contaminated with experimental noise due to biases, which makes proper normalization procedures crucial in processing WES data. Many methods have then been developed to remove biases and restore the true CNV signals, such as VarScan2 (Koboldt *et al.*, 2012), CLAMMS (Packer *et al.*, 2016), CoNIFER (Krumm, Peter H. Sudmant, *et al.*, 2012), EXCAVATOR (Magi *et al.*, 2013), and CODEX (Jiang *et al.*, 2015) (see Section 1.2.2 for a comprehensive review). Among them, two normalization approaches have been commonly used in the existing literature that is relevant to our study, including a median normalization approach and a Poisson latent factor (PLF) model-based approach. Median normalization approach used a three-step

procedure to sequentially remove three observed sources of bias (i.e., exon length, GC-content, and mappability) (Magi *et al.*, 2012). Specifically, for each exon, it corrects raw WES data according to the deviation of the median intensities of exons that have the same biases values as this exon from the overall median. As discussed in Section 1.2.2, median normalization approach is effective and computationally efficient in removing known sources of bias, especially when processing large-scale datasets. This median approach was implemented in tools EXCAVATOR (Magi *et al.*, 2013) and EXCAVATOR2 (D’Aurizio *et al.*, 2016). Recently, CODEX (Jiang *et al.*, 2015) and CODEX2 (Jiang *et al.*, 2018) used a PLF model-based method, which is capable of removing both systematic and observed biases. The proposed PLF model includes several linearly additive terms to model the observable biases (i.e., exon length, mappability score, and GC content). In addition, it also includes terms in the form of the latent factors, which can specifically capture unobservable biases due to other unknown experimental variables. Following the data normalization step, statistical segmentation algorithms are applied to locate all the change points and identify the segments that share the same underlying copy number. Most WES CNV detection tools still use algorithms that were previously designed to analyze SNP array data (see detailed reviews in Section 1.2.1 and 1.2.2). For example, ExomeDepth (Plagnol *et al.*, 2012), CoNIFER (Krumm, Peter H Sudmant, *et al.*, 2012) andXHMM (Fromer *et al.*, 2012) adopted HMM as their segmentation algorithm. On the other hand, ExomeCNV (Sathirapongsasuti *et al.*, 2011), CODEX (Jiang *et al.*, 2015), and CODEX2 (Jiang *et al.*, 2018) all use CBS algorithm. Still, these methods assume that observed data across different loci are independent. In Chapter 2, we have demonstrated the correlation structure in SNP array data which

demonstrated the necessity of integrating the non-independence structure in segmentation algorithms. However, it is yet to be determined whether the correlation structure systematically exists in WES data, especially after applying different normalization methods. Moreover, given different correlation structures of the normalized data generated from different normalization methods, it remains uncertain about the necessity of integrating correlation structure in profiling WES data and how such correlation structure integration can improve the accuracy of CNV detection in statistical modeling.

In this chapter, we first conducted an initial evaluation to explore the correlation structure in WES data before and after median normalization and PLF normalization. It was found that the median normalization approach retained the correlation structure of raw read count data, while the Poisson latent factor (PLF) based approach significantly reduced the correlation. Based on such findings, we proposed a correlation-based method, CORRseq, which utilized the theory from LDcnv segmentation algorithm to systematically integrate the correlation structure in median normalized data. Due to the weak correlation structure of PLF normalized data, we also developed an alternative independent-based method in parallel, so as referred to SARAsq, which was built upon on our previously developed SNP array method modSaRa (Xiao *et al.*, 2017). Spike-in simulation studies and a further application to a real WES dataset with experimentally validated CNVs showed that CORRseq significantly improved the detection accuracy, especially for that of medium and large CNVs. CORRseq was further applied to the international TRICL dataset to identify lung cancer-associated CNVs. Methods developed in this chapter were included in a comprehensive and user-friendly R package, “CORRseq” (<https://github.com/adamluo12/CORRseq>).

3.3 METHODS

3.3.1 EVALUATION OF CORRELATION STRUCTURE IN REAL WES DATA

To explore the genomic correlation structure in the raw read counts of WES data, we utilized data of 89 healthy samples from the 1000 Genomes Project (Auton, Gonalo R Abecasis, *et al.*, 2015). Among them, 46 samples were sequenced at the Baylor College of Medicine using the Illumina HiSeq 2000 and 2500 platforms with the Roche HGSC VCRome capture kit and 43 samples were sequenced at the Washington University using the Illumina HiSeq 2000 platform with the Nimblegen SeqCap EZ Human Exome Library v.3.0 capture kit. The detailed experimental samples and genotyping information was described in previous literature (Jiang *et al.*, 2018). Five normal samples from the 1000 Genomes project were arbitrarily chosen and treated as control samples in calculating the genetic intensities with the normalization procedures.

With these samples, the correlation structure was evaluated before and after two commonly used normalization procedures: a three-step median normalization method (Magi *et al.*, 2013) and a PLF-based normalization method (Jiang *et al.*, 2015). The median normalization approach used the exon mean read count (EMRC), which was defined as $EMRC = RC/L$ for each exon region, where RC represented the number of reads mapped to the exon region, and L was the exon length. A three-step procedure was then implemented to remove the effects from the GC content, mappability score and exon size by $\overline{EMRC}_i = EMRC_i \times \frac{m}{m_e}$, where $EMRC_i$ was the mean read count for exon i , m_e was the median EMRC of all the exons with the same e value (where $e = [\text{GC content, mappability score, exon size}]$) as the i -th exon, and m was the overall median of all the exons. The procedure was used to normalize both testing and control samples. Control

samples were then pooled together, across which the mean EMRC was calculated by exon and used as the reference. Logarithm transformation of the ratio of EMRC from the testing sample and that of the pooled reference was calculated (i.e., *log2R-MED*) and used as input for CNV detection. Meanwhile, the PLF-based normalization method fitted a Poisson latent factor model to simultaneously correct the measurable sources of biases and unmeasurable systemic biases. Afterward, an iterative maximum-likelihood algorithm was adopted to estimate the expected read counts under the null condition (i.e., there is no CNV). Logarithm transformation of the ratio of observed read counts and expected “null” read counts were calculated and referred to as *log2R-PLF*.

For the evaluation of correlation structure, 200 consecutive exons were randomly chosen from chromosomes 1, 5, and 9 and were studied in all samples. We investigated the correlation structure of the raw read depth, *log2R-MED*, and *log2R-PLF* by calculating the overall strength of the correlation represented by the average squared correlation coefficient estimates, $\overline{r^2}$.

3.3.2 CORRseq AND SARaseq

Given the obvious correlation structure observed in the *log2R-MED* data (shown in Figure 3.2B), we developed CORRseq, which systematically integrated the data correlation structure in a change point detection model for CNV profiling with WES data. Specifically, let $\mathbf{Y} = (Y_1, \dots, Y_M)^T$ denote the normalized *log2R-MED* on M exons in a single sequence for each sample, we still assume the genetic intensities follow a multivariate normal distribution given the dependence structure of the genome as defined in Section 2.3.2. In parallel, as the PLF-based normalization approach dramatically eliminates the correlation structure (as shown in results with Figure 3.2C), the SARaseq

method was developed, which instead assumed the independent observations. As a result, the main difference of the segmentation algorithm with CORRseq and SARAsseq methods mainly relies on the assumption of the observed data.

Figure 3.1 illustrated the workflows of our proposed CORRseq and SARAsseq methods. CORRseq requires WES read count data from both testing and control samples as inputs. Read counts are normalized using a median normalization approach and segmented by a correlation-based algorithm described in Section 2.2.2-2.2.3 (i.e., LDcnv). SARAsseq only requires WES read count data from testing samples. WES read counts are then normalized using a PLF-based normalization approach and analyzed by an independent assumption based algorithm (modSaRa). SARAsseq implement the same procedures as CORRseq, except using a different distribution of $D(x)$, $D(x) \sim N\left(0, \frac{2\sigma^2}{w}\right)$.

3.3.3 SPIKE-IN SIMULATION AND APPLICATION TO WES DATA FROM THE 1000 GENOMES PROJECT

To evaluate the performance of our proposed methods (i.e., CORRseq and SARAsseq), we conducted in silico spike-in studies that best retained the noise and correlation structure of the real data. With the raw read depth data from chromosomes 1 and 2 in the 70 samples from the 1000 Genomes Project (Auton, Gonçalo R Abecasis, *et al.*, 2015), we first applied filters to remove the exons across all samples that may contain CNVs. Specifically, the filtering step excluded exons those were either detected by our methods (CORRseq and SARAsseq), or two commonly used methods, CODEX2 (Jiang *et al.*, 2018) and EXCAVATOR2 (D’Aurizio *et al.*, 2016), or those were reported by the Database of Genomic Variants (MacDonald *et al.*, 2014). As such, the remaining sequences were treated as the background CNV-free sequences. This background CNV-

free data was also used as negative controls for median and PLF normalization procedures. Second, for each sample, signals of 60 CNV segments were added to the background data with varied lengths (short: 5~20 exons, medium: 20~50 exons, and long: 50~80 exons) and varied copy number states (deletion and duplication), respectively. Deletions were spiked in by multiplying the background depth of coverage by a normal random variable $N(0.2, 0.1^2)$, and duplications were spiked-in similarly with another normal random variable $N(5.0, 0.1^2)$ (Jiang *et al.*, 2018).

With the simulated data, we evaluated the performance of the proposed CORRSeq by benchmarking to EXCAVATOR2 and CODEX2. EXCAVATOR2 uses a Heterogeneous Shift Level Mean (HSLM) model (Magi *et al.*, 2010, 2011) for segmentation, and a FastCall algorithm (Benelli *et al.*, 2010) was utilized for CNV classification. A CBS algorithm (Olshen *et al.*, 2004) was used for the segmentation in CODEX2 (Jiang *et al.*, 2018), assuming independent observations. These methods were assessed by using the precision rate $\left(\frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}\right)$, recall rate $\left(\frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}\right)$, and $F1$ score $\left(2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}\right)$.

We also evaluated the performance of our methods by analyzing 36 healthy individuals from the 1000 Genomes Project (Auton, Gonalo R Abecasis, *et al.*, 2015). CNV calling was assessed using the “validation sets” from three high-quality microarray CNV studies (International HapMap 3 Consortium, 2010 (D. Altshuler *et al.*, 2010); Conrad *et al.*, 2010 (Conrad *et al.*, 2010); McCarroll *et al.*, 2008 (McCarroll *et al.*, 2008)). These three validation datasets have samples being genotyped by both SNP array and WES therefore serve as gold standard datasets for many existing studies (Xiao *et al.*, 2019; Luo *et al.*, 2020; Jiang *et al.*, 2018). In total, there were 1,327,6,700 and 1,946

high-quality CNVs from these three datasets, respectively. More details of these validation sets can be found at Luo *et al.* (Luo *et al.*, 2020). To improve the quality of “validation sets” for CNV evaluation, stringent quality control procedures were conducted (i.e., the reported CNV must overlap with two exons, have less 5% missing rate across all samples). With these three validation sets, our methods were compared to CODEX2 and EXCAVATOR2. Still, for EXCAVATOR2 and CORRseq, additional five samples from the 1000 Genomes Project were selected and used for control samples in the normalization step. Similarly, each method was evaluated using the precision rates, recall rates, and *F1* scores.

3.3.4 ANALYSIS OF TRICL CASES AND CONTROLS

We applied CORRseq and EXCAVATOR2 to the Transdisciplinary Research Into Cancer of the Lung (TRICL) with 1,084 cases and 919 controls (Amos *et al.*, 2017). Study samples were sequenced at the Center for Inherited Disease Research (CIDR) using Illumina HiSeq2500 platform. the raw sequence data (FASTQ format) was processed via GATK best practice workflow to produce BAM files by using 1000 genomes phase 2 reference (Auton, Gonçalo R Abecasis, *et al.*, 2015; Geraldine A. Van der Auwera, 2020). Data clean-up procedures such as base call quality recalibration variant filtering and genotypes refinement were performed. Principal component analysis (PCA) was performed on quality metric generated from sequencing pipeline to exclude quality outliers. Kinship coefficient between each pair of participants was also calculated to identify and exclude duplicated and related samples.

To keep high-quality CNV calls for the downstream association analysis, post-calling quality control (QC) filters were applied. Specifically, the filter step remove

individual CNV calls that (1) are overlapped with centromeric regions; (2) are extremely short (i.e., < 5 exons) or long (i.e., >80 exons for CORRseq and >200 exons for EXCAVATOR2). Individual samples were excluded if they had >100 CNVs. Finally, CNV calls from these two methods were combined as the final calling set.

To identify CNV loci that confer risk for lung cancer, logistic regression was performed to test CNV association at both the level of individual genes and CNV regions (CNVRs),

$$\text{logit}(P(\text{disease} = 1)) = \beta_0 + \beta_{del}DEL + \beta_{dup}DUP + \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\beta}' \sum_{i=1}^4 PC_i \quad (9)$$

CNVRs were constructed using tool CNVRuler by merging overlapped CNV segments (i.e., at least 1 bp) and trimming any long, rare regions (Kim *et al.*, 2012). *DEL* and *DUP* were two indicator variables for deletions and duplications separately. We adjusted the covariates including smoking status (ever/never), age, and gender. Top 4 principal components were also included to adjust for the population structure (Patterson *et al.*, 2006).

3.4 RESULTS

3.4.1 CORRELATIONS STRUCTURE EXISTS IN WES DATA FROM THE 1000 GENOMES PROJECT

To understand the correlation structure in WES data and the potential influence from normalization steps, we first utilized the 1000 Genomes project data to investigate the correlation pattern in WES data. The correlation structure of the raw read depth, and that of normalized by median and PLF-based normalization approaches were evaluated separately. As a result, a moderate correlation ($\overline{r^2} = 0.38$) among exons was observed in

both WES raw read counts and median normalized data (Figure 3.2A-B). On the contrary, the latent factor-based normalization approach dramatically reduced the underlying correlation structure ($\overline{r^2} = 0.05$) (Figure 3.2C), as most of the covariance captured by top latent factors were removed during normalization procedure (Friguet *et al.*, 2009). Together, such result provided strong evidence of the existence of correlation structure among exons in raw WES data and the median normalized data, which motivated the development of CORRseq as described in Section 3.2.2. Given the independent structure of the PLF normalized dataset, we also developed SARaseq method for handling such non-correlated data.

3.4.2 EVALUATION OF CORRseq VIA SPIKE-IN STUDIES AND APPLICATION TO THE 1000 GENOMES PROJECT

We first evaluated the performance of our methods using spike-in simulations under different scenarios. When the CNV size was greater than 20 exons, CORRseq performed the best in almost all the scenarios. For example, for duplications with length 20~50 exons, CORRseq had an $F1$ score of 97.23, while the corresponding $F1$ score was 92.62 for EXCAVATOR2, and 67.53 for CODEX2. Besides, the superior performance of the independence structure-based method, SARaseq, was clearly observed in detecting short CNVs (i.e., 5~20 exons) (Figure 3.3). For example, for deletions with length 5~20 exons, SARaseq had an $F1$ score of 87.41, while the corresponding $F1$ score was 83.52 for CORRseq, 81.52 for EXCAVATOR2, and 67.98 for CODEX2, respectively (Table 3.1). It was also noteworthy to mention that both SARaseq and CORRseq outperformed CODEX2 and EXCAVATOR2 in almost all the scenarios, except for detecting deletions with CNV length 50~80 exons. CODEX2 presented nearly perfect recall rates in all

scenarios, but the precision rates were compromised. This is because CODEX2 implemented CBS-based procedure, which tended to be sensitive and may capture many false positive signals (Niu and Zhang, 2012).

We further applied our methods to a real data benchmarking against EXCAVATOR2 and CODEX2. Three high-quality CNV validation datasets from HapMap3 (D. Altshuler *et al.*, 2010), Conrad *et al.* (Conrad *et al.*, 2010), and McCarroll *et al.* (McCarroll *et al.*, 2008) were utilized for evaluation. Among them, the majority of the “true” CNVs were short CNVs (2~20 exons) (i.e., 91.9%, 90.7% and 88.6% for HapMap3, Conrad *et al.*, and McCarroll *et al.*, respectively). The overall performance of SARaseq was better than that of other three methods in HapMap3 and Conrad *et al.* datasets with the highest *F1* scores (Table 3.2, Figure 3.4). In the Conrad *et al.* dataset, SARaseq had an *F1* score of 24.66, while the corresponding *F1* score was 17.50 for CORRseq, 22.57 for CODEX2, and 22.92 for EXCAVATOR2 (Table 3.2). EXCAVATOR2 achieved high recall rates but suffered from low precision. These results further demonstrated the advantage of SARaseq in detecting short CNVs.

In conclusion, our proposed CORRseq method, which integrated the correlation information, presented significantly improved performance in CNV detection, especially for the detection of medium and large CNVs. Meanwhile, the independence assumption based SARaseq presented a consistent power gain in detecting short CNVs. These results further suggested that we need to design statistical methods in the segmentation step by evaluating the correlation structure of the data after normalization.

3.4.3 APPLICATION TO TRICL STUDY

A total 1,033 lung cancer cases and 875 control samples passed the pre-calling QC (Supplementary Table A.7). Using the TRICL WES datasets, we identified 166,062 CNVs in autosomes from two methods in total. Overall, we detected more deletions than duplication (113,093 vs. 52,969). No significant difference in the overall proportion of deletions and duplications between cases and controls was observed (49% vs. 51% for deletions and 52% vs. 48% for duplications, respectively). Also, there was no difference in the length of deletions and duplications between cases and controls (Supplementary Table A.8).

Those identified CNVs were mapped to 4,992 genes and 4,469 CNVRs. Those CNVRs included 147 common CNVRs (i.e., occurring in $\geq 1\%$ of samples), 2,036 rare CNVRs (i.e., $<1\%$), and 2,286 singletons. Our gene-based association analysis identified the deletion genes *FCGR3A* in 1q23.3 region (OR=1.76, 95% CI=1.28-3.13, P-value=0.002), *AMBRA1* in the 11q11.2 region (OR=3.71, 95% CI=1.40-9.83, P-value=0.008), *NPEPPS* in the 17q21.32 region (OR=1.37, 95% CI=1.08-1.75, P-value=0.01) and duplication genes *ZNF280A* in the 22q11.22 region (OR=1.4, 95% CI=1.14-1.70), *PRAME* in the 22q11.22 region (OR=1.37, 95% CI=1.12-1.68, P-value=0.001), *AMY2B* in the 1q21.1 region (OR=1.80, 95% CI=1.14-2.82, P-value=0.01) (Supplementary Table A.9). Using CNVR-based association testing, results highlighted the deletion regions in 1p36.33, 12q23.3, 17q21.32, 10q25.3, 19p13.3, 3p14.3 and duplication regions in 22q11.22, 13q14.11, 12p13.2 (Supplementary Table A.10).

3.4.4 PACKAGE DEVELOPMENT AND EXAMPLE USAGE

Here, we build a user-friendly R package, CORRseq, designed to perform the CNV detection with WES data using the proposed CORRseq and SARAsseq method. The package includes procedures for mapped read count calculation, data normalization, and CNV calling. Below, we present a real data example to illustrate the usage of the package.

CORRseq requires two types of input files: genome sequencing data files (.BAM) and WES target file (.BED). Function CODEX2::getcoverage() is used to generate raw read count data matrix. The output data matrix contains information for one exon per row, and each column represents data for one study sample. CORRseq also accepts direct input of read count data generated by using other tools, such as SAMTools and GATK (Li *et al.*, 2009; Geraldine A. Van der Auwera, 2020). CORRseq::normalize() is then applied to perform median normalization approach and generate *log2R-MED* for CNV calling. An optional smoothing step, CORRseq::smooth() is recommended to remove the potential outliers of *log2R-MED*. After that, CORRseq::CNVout() is implemented to generate CNV calls. CORRseq produces comma delimited files (.csv) that has 8 columns with each row corresponding to one CNV records. The first four columns contain sample ID, chromosome, start and end positions of CNV in base pair. While the last three columns contain detailed information of the identified CNV segments, including length of the CNV in base pair as well as in number of exons, and the copy number state of the called segment (deletion or duplication). In parallel, we also developed procedures of performing SARAsseq, which uses CORRseq::PLF() for normalization and CORRseq::SARAsseqCNVout() for CNV identification.

Figure 3.6 displays the plots of $\log_2 R\text{-}MED$ of four CNVs on chromosome 6 identified by CORRseq using WES data from the 1000 Genomes Project (Auton, Gonalo R. Abecasis, *et al.*, 2015). In the absence of CNV events, $\log_2 R\text{-}MEDs$ are expected to be clustered around zero. Figure 3.6a shows two typical examples of deletions, where the signal intensities within the detected region are below zero. While intensities within two identified regions in figure 3.6b are above zero, which indicate the presence of duplication events.

3.5 DISCUSSION

In this chapter, we developed a correlation-based method CORRseq, built on our previous SNP array method with the capability of analyzing WES data coupled with the median normalization procedure. We also developed SARaseq utilizing an independent assumption-based segmentation algorithm and the PLF-based normalization procedure. Through simulations and applications, we demonstrated the desirable performance of these methods and their application scopes.

Our preliminary study explored the correlation structure in the real data and investigated the potential influence from different normalization methods on the design of segmentation methods. This is the first report to evaluate the impact of different normalization methods on the correlation structure of WES CNV. We found that latent factor-based normalization method significantly reduced the correlation structure inherent in the read count data, while median normalization method retained the correlation structure. This finding provides great direction for developing CNV detection methods in future studies. Intuitively, we may develop independence assumption-based segmentation method with PLF-based normalization and correlation-based segmentation methods

coupled with median normalization, though more real data shall be studied and exhibited in the showcase. Also, the different correlation structure presented in the real data after median and PLF-based normalization led us to an inspiring question, shall we consider the genomic correlation structure in developing statistical methods for CNV detection? Indeed, similar concerns have been dabbled in previous literature but have not been thoroughly discussed. Fromer *et al.* and D'Aurizio *et al.* both assumed the adjacent exon targets in genome were correlated, and the correlation strength was directly related to the genomic distance between exons (Fromer *et al.*, 2012; D'Aurizio *et al.*, 2016). Zhang *et al.* and McCarroll *et al.* analyzed the common CNVs and employed the assumption that the correlations structure exist between chromosomal sites within CNV regions (Q. Zhang *et al.*, 2010; McCarroll *et al.*, 2008). Moreover, Wei *et al.* recommended that the CNV detection methods should take the correlation of read depths into consideration in analyzing next-generation-sequencing data (Wei and Huang, 2020). In this study, our results supported this evidence and suggested that segmentation methods without considering correlation structure can identify short CNVs with high accuracy though they need to be coupled with normalization procedure that removes the correlation. Also, the correlation-based segmentation method coupled with normalization that retains the correlation would perform better for detecting medium and long CNVs. These findings imply the advantage of utilizing genomic correlation information on detecting medium and long CNVs with improved accuracy though it gives little gains for small-sized CNVs.

There are still limitations in our study. First, our methods adopted a single sample scanning strategy, which may result in high false discovery rate and low sensitivity due to sample-specific noise and CNV complexity. Multi-sample based CNV detection methods

that borrow the comprehensive information from multiple samples is thought to improve the robustness and detection power with noisy data (Song *et al.*, 2016; Wang *et al.*, 2020). Our methods have potential to be extended to the multi-sample setting, where the main challenge is developing scan statistics that can effectively integrate information across samples. Second, our methods merely exploited the WES read counts from exon regions, limiting the possibility to study the CNVs on non-coding regions. While in many of these WES cohorts, the SNP array data for the same samples are also available. Methods that can integrate information from multiple platforms (i.e., WES and SNP array) is expected to offer a full-coverage CNV detection. Therefore, in the following chapter, we focus on developing statistical approach that can effectively integrate multi-platform and multi-sample data.

TABLES AND FIGURES

Table 3.1 Assessment of CORRseq and SARAsq based on simulation data. CNV calls were generated by CORRseq, SARAsq, CODEX2, and EXCAVATOR2. The CNV size varied from 5~20, 20~50 to 50~80 exons. Precision rates and recall rates, as well as the F1 scores, were summarized for each method.

CNV Length (exon)	Methods	Deletions (%)			Duplications (%)		
		Precision	Recall	F1	Precision	Recall	F1
5~20	CORRseq	89.16	78.55	83.52	90.22	81.48	85.62
	SARAsq	95.72	80.43	87.41	83.89	97.05	89.99
	CODEX2	51.51	99.95	67.98	45.99	100.00	63.01
	EXCAVATOR2	78.17	85.17	81.52	78.41	65.36	71.29
20~50	CORRseq	90.02	99.00	94.30	94.66	99.95	97.23
	SARAsq	91.31	88.57	89.92	92.23	99.43	95.69
	CODEX2	53.37	100.00	69.59	50.98	100.00	67.53
	EXCAVATOR2	80.34	98.93	88.67	86.72	99.38	92.62
50~80	CORRseq	93.14	99.17	96.06	89.63	100.00	94.53
	SARAsq	71.79	93.57	81.25	90.66	99.12	94.70
	CODEX2	46.84	100.00	63.80	59.26	100.00	74.42
	EXCAVATOR2	79.97	99.36	88.62	84.65	99.69	91.56

Table 3.2 Assessment of CORRseq and SARAsq based on the 1000 Genomes Project data. Assessment of CNV calls generated by CORRseq, SARAsq, CODEX2, and EXCAVATOR2 using validation sets from HapMap3, Conrad *et al.*, and McCarroll *et al.*. Precision rates and recall rates, as well as the *F1* scores, were summarized for each method.

Methods	HapMap3 (%)			Conrad (%)			McCarroll (%)		
	Precision	Recall	<i>F1</i>	Precision	Recall	<i>F1</i>	Precision	Recall	<i>F1</i>
CORRseq	17.45	40.62	24.41	12.12	31.46	17.50	14.15	37.04	20.48
SARAsq	48.39	67.57	56.39	19.42	33.79	24.66	36.52	56.03	44.22
CODEX2	45.37	56.32	50.26	22.40	22.75	22.57	45.69	50.48	47.97
EXCAVATOR2	17.10	64.08	27.00	14.64	52.79	22.92	19.03	69.09	29.84

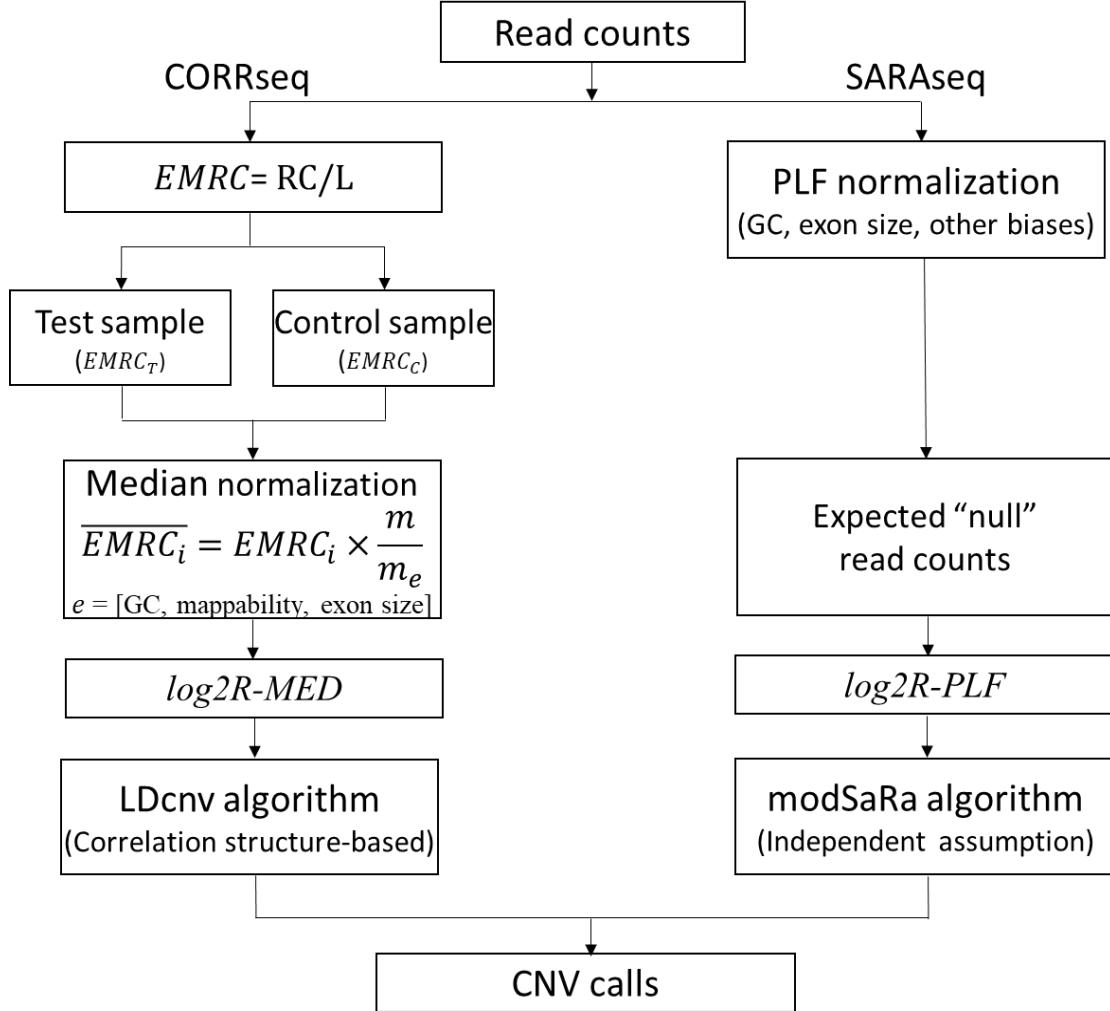


Figure 3.1 Analysis workflow of CORRseq and SARaseq. CORRseq requires WES read count data from both testing and control samples as inputs. Read counts are normalized using a median normalization approach to adjust exon length, GC-content, and mappability biases. Logarithm transformation of the ratios of normalized data from the testing and the pooled control samples is calculated, which is referred to as *log2R-MED*. CORRseq implements a correlation-based algorithm (LDcnv) to identify CNVs. SARaseq only requires WES read count data from testing samples. Read counts are normalized using a PLF-based approach to mitigate observable biases (GC content, amplification efficiency, and exon size) and latent systemic biases and estimate the expected “null” read counts. Logarithm of the observed read counts and expected “nulls” are calculated, which is referred to as *log2R-PLF*. SARaseq uses an algorithm assuming independence (modSaRa) for the CNV identification.

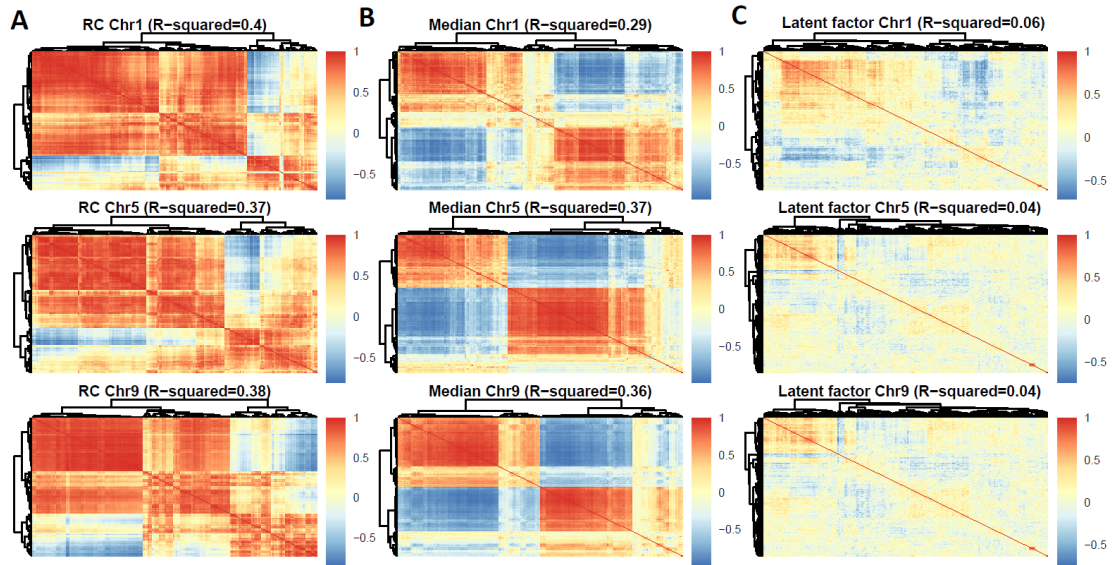


Figure 3.2 Assessment of genomic correlation structure of WES data from 1000 Genomes project. The left panel is the correlation heatmaps across 200 consecutive exons in chromosome 1, 5, and 9 for WES raw read counts (RC) data (A). The middle and right panels are the correlation heatmaps of WES data after median normalization (B) and latent factor-based normalization (C). Median normalization retains the correlation structure of raw read counts, while the PLF-based approach significantly reduces the underlying correlation.

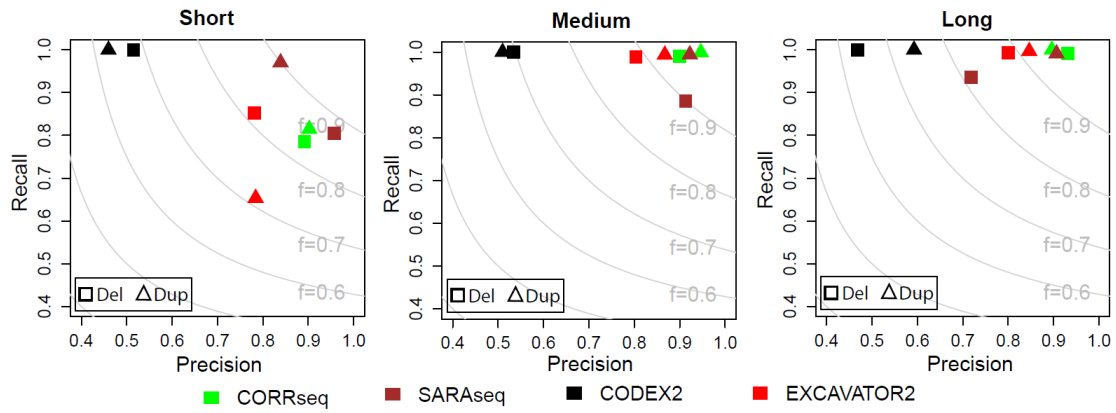


Figure 3.3 Assessment of CORRseq and SARaseq based on simulation data. CNV calls were generated by CORRseq, SARaseq, CODEX2, and EXCAVATOR2. The CNV size varied from 5~20 (short), 20~50 (medium) to 50~80 (long) exons. Precision rates and recall rates, as well as the $F1$ scores, were summarized for each method. Del: deletion; Dup: duplication.

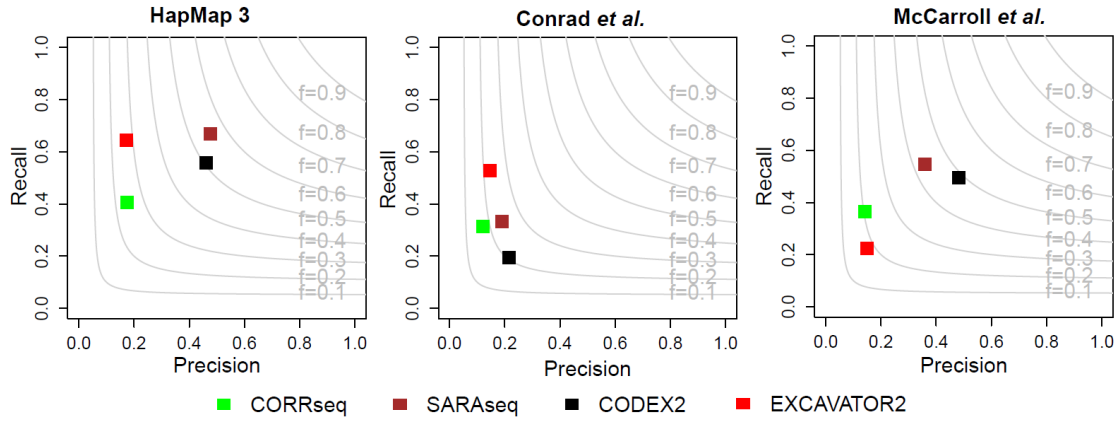


Figure 3.4 Assessment of CORRseq and SARaseq based on the 1000 Genomes Project data. Assessment of CNV calls generated by CORRseq, SARaseq, CODEX2, and EXCAVATOR2 using validation sets from HapMap3, Conrad *et al.*, and McCarroll *et al.*. Precision rates and recall rates, as well as the *F1* scores were summarized for each method.

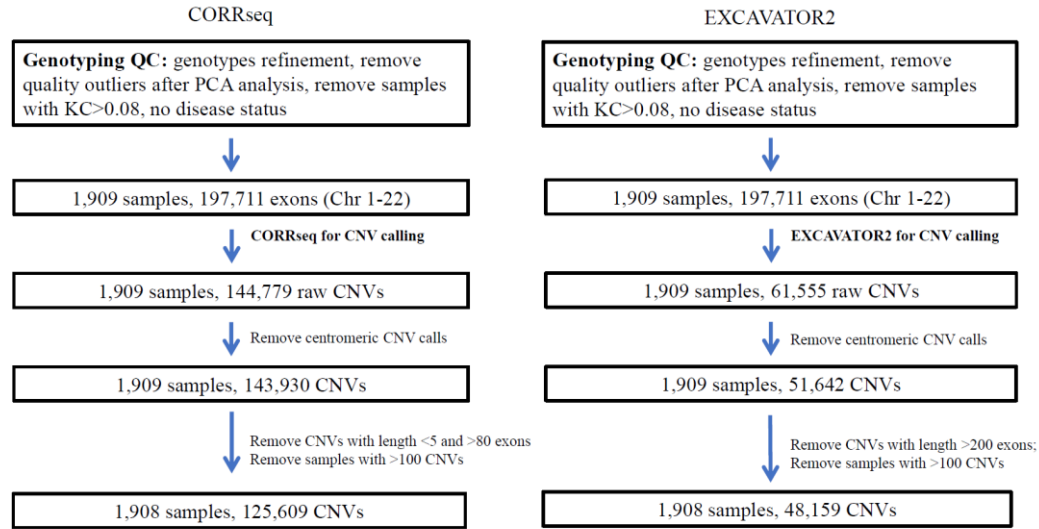


Figure 3.5 Overview of the application to the TRICL case-control study. The figure outlines the study design with a brief description of quality control (QC) steps. Summary of key results includes the sample size and number of CNVs at various stages of analysis. PCA: principal component analysis; KC: kinship coefficient; Chr: chromosome.

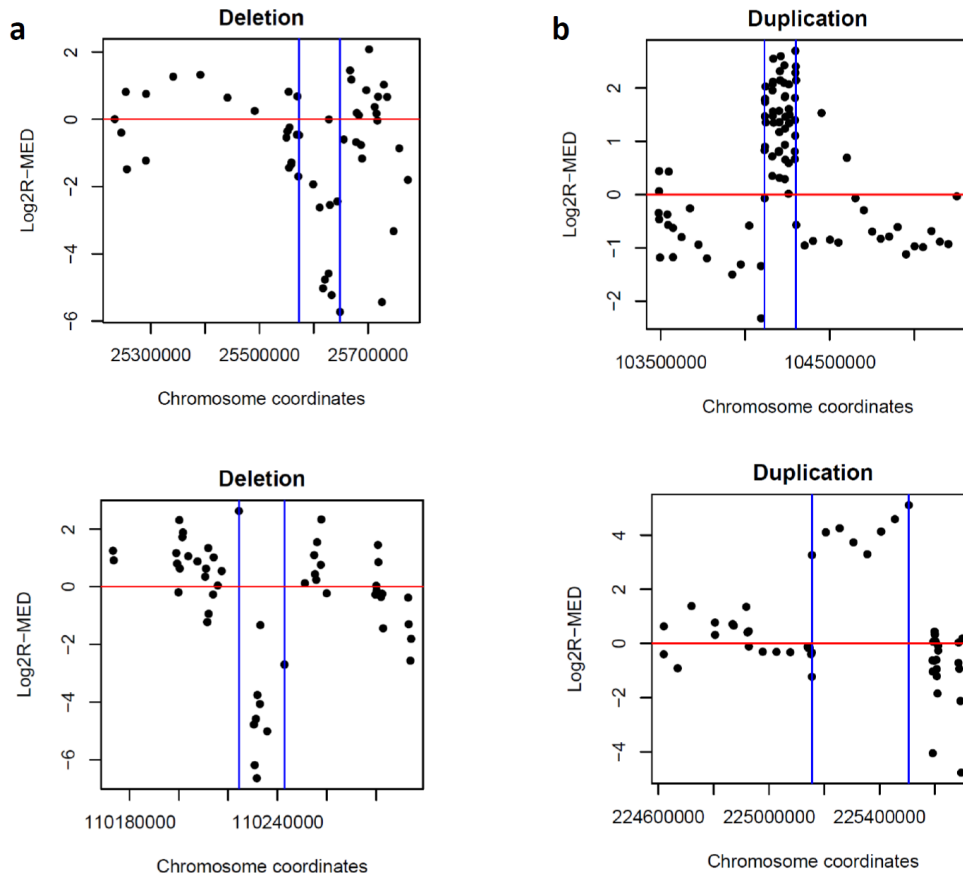


Figure 3.6 CNVs identified by CORRseq from the 1000 Genomes Project. The figure shows the data intensities of four CNVs (two deletions and two duplications). The regions between two blue vertical lines are identified CNVs. X-axis indicates the genomic position on chromosome 6 and the y-axis is the signal intensity (i.e., Log2R-MED). Column (a): deletions; (b) duplications.

CHAPTER 4

BMI-CNV: A BAYESIAN FRAMEWORK FOR MULTIPLE GENOTYPING PLATFORMS DETECTION OF COPY NUMBER VARIANTS³

4.1 ABSTRACT

Whole-exome sequencing (WES) enables detection of CNVs with high resolution in functional protein-coding regions. However, variations in the intergenic or intragenic regions are excluded from these studies. Fortunately, in many existing large cohorts, samples have been previously sequenced by different genotyping platforms, such as SNP array. As a result, methods for integrating multiple genotyping platforms are highly demanded for improved CNV detection. Moreover, conventional single sample-based CNV calling methods often suffer from high false discovery rate. A multi-sample strategy may reduce detection error and will be more robust to data variations.

We developed BMI-CNV, a Bayesian Multi-sample and Integrative CNV (BMI-CNV) profiling method with data sequenced by both WES and microarray. By incorporating complementary and concurrent information from multiple platforms, our method can accurately detect CNVs with a genome-wide scale. With extensive simulations, BMI-CNV outperformed existing methods with remarkably improved accuracy for both multiple and single platform analyses. By applying to the matched 1000 genomes project and HapMap project data, we showed that BMI-CNV accurately

³ Xizhi Luo, Guoshuai Cai, Alexander C. McLain, Christopher I. Amos, Bo Cai, Feifei Xiao. Submitted to *Bioinformatics*, 06/24/2021.

detected common variants. We further applied it to the TRICL consortium with matched WES and OncoArray data and generated preliminary results involving several lung cancer risk associated genes in 17q11.2, 1p36.12, 8q23.1 and 5q22.2 regions, which may provide new insights into the etiology of lung cancer.

4.2 INTRODUCTION

With the dramatic growth of modern technologies and the accompanying cost drop in sequencing, massive WES datasets have been generated from large-scale biomedical studies, which allows for the identification of genomic variants in functional protein-coding regions (Amos *et al.*, 2017). However, exons only encompass 1% of the genome, limiting the possibility to investigate the impact of CNVs located in the non-coding regions (Craig Venter *et al.*, 2001). Moreover, WES is subject to the non-uniform coverage of sequence reads in the assembly procedure due to the existence of short duplications or deletions, resulting in many dropped out segments which are originally mapped to the exome. For example, Fang *et al.*, found that more than 16% of the exons cannot be captured by WES experiments (Fang *et al.*, 2014). This will lead to negligence in detecting short CNVs and therefore integrating available SNP array has great potential to overcome this challenge. In many of these WES cohorts, the same samples have been previously genotyped by the SNP array. For instance, the international Transdisciplinary Research In Cancer of the Lung (TRICL) consortium genotyped 2,003 subjects with both WES and SNP array data (Amos *et al.*, 2017). The Alzheimer's Disease Genetics Consortium (ADGC) and Alzheimer's Disease Sequencing Project (ADSP) (Karch *et al.*, 2016; Beecham *et al.*, 2017) have also collected such multi-platform data. Consequently,

the demand for multi-platform (e.g., WES and SNP) integration methods, which will comprehensively study CNV in a full-coverage manner, has dramatically increased.

Similar efforts such as iCNV have been made by Zhou *et al.* for integrative segmentation (Zhou *et al.*, 2018). In iCNV, data from different platforms were first normalized and standardized and then jointly segmented by a Hidden Markov Model (Zhou *et al.*, 2018). This method had a significant boost in accuracy compared to WES, however, it only used information from a single sample. As we know, technological and biological factors prominent in real data usually increase the variations and noise in data intensities, leading to unreliable findings with single-sample scanning of CNVs. Consequently, multiple sample strategies previously introduced for detecting common CNVs can improve the robustness and detection power with noisy data. Such a direction has been supported by various existing studies (N. R. Zhang *et al.*, 2010; Siegmund *et al.*, 2011; Song *et al.*, 2016), but none has focused on multi-platform integration. Moreover, most widely used WES methods such as CODEX2 and EXCAVATOR are also for single-sample scanning (Magi *et al.*, 2013; Jiang *et al.*, 2018). As a result, it is of highly demand to develop a full spectrum CNV detection method that can meanwhile achieve high accuracy using comprehensive information from all samples.

In this study, we developed BMI-CNV, a Bayesian Multi-sample and Integrative CNV calling method. Comparing to existing approaches, BMI-CNV efficiently integrates data from multiple platforms (i.e., WES and SNP array) and multiple samples to exploit the comprehensive information, leading to a full spectrum study of CNVs. Extensive numerical simulation studies showed that BMI-CNV presented significantly improved sensitivity over the existing methods. The method was further illustrated by applying to

the HapMap project (D. M. Altshuler *et al.*, 2010) and the 1000 Genomes project (Auton, Gonçalo R. Abecasis, *et al.*, 2015). It was further applied to the international TRICL dataset to identify lung cancer-associated CNVs. This new method has a wide scope of applications and has great potential to be further extended to profile CNVs for whole-genome sequencing and single-cell sequencing data analyses.

4.3 METHODS

METHODS OVERVIEW

Our method mainly focuses on CNV detection by integrating the SNP array and WES data, although it can also be naturally applied to the WES data only situation. Figure 4.1 shows an overview of the framework of BMI-CNV. First, WES read counts and SNP array intensities are integrated using a series of data integration procedures, including normalization, standardization, and merging. Our main algorithm consists of two main stages: Stage I uses a Bayesian PSBP method (Section 4.2.2) coupled with a Gaussian mixture model-based initial data filtering (Section 4.2.4) to identify shared CNV regions, and Stage II as the individual CNV calling procedure (Section 4.2.3).

4.3.1 DATA DESCRIPTION AND MODELS

First, we performed platform-specific normalization procedures for the original data. For WES data, we utilized the EXCAVATOR2 median normalization procedure to generate *log2R-MED* as described in section 3.3.1, with external controls being pre-specified by researchers (D’Aurizio *et al.*, 2016). This *log2R-MED* data was then processed by the lowess-scatter plot procedure to adjust read-depth differences between testing and control samples and remove coverage dependent bias. For array data, PennCNV was used to

adjust the genomic wave on genetic intensities (i.e., Log R ratio (LRR)) (Wang *et al.*, 2007).

To bring the SNP array LRR values and the WES-derived *log2R-MED* to the same scale, we standardized each data via a robust scaling approach (Rousseeuw and Croux, 1993). Compared to the conventional standardization method, the robust scaling approach used median and interquartile ranges to mitigate the influence from potential outliers and signals from double deletions (Details in Supplementary C.1). The WES and SNP array data were merged by chromosomal coordinates to effectively integrate information for joint segmentation.

Let Y denote a $n \times m$ data matrix obtained from the pre-CNV calling procedures described above, where Y_{ij} represented the processed genetic intensities (e.g., LRR from array or *log2R-MED* from WES) for the i -th ($i=1, 2, \dots, m$) marker (e.g., SNPs from array or exons from WES) in sample j ($j=1, 2, \dots, n$). We assumed a classic normal kernel for Y_{ij} ,

$$Y_{ij} \sim N(Y_{ij} | \phi_i) \quad (10)$$

$\phi_i = (\mu_i, \sigma_i^2)$ is an unknown vector of the underlying mean and variance at position i across all samples, in which different values of ϕ_i indicated the existence of different copy number states. We assumed there were five copy number states, including the deletion of a single copy (del.S), deletion of double copies (del.D), diploid, duplication of a single copy (dup.S), and duplication of double copies (dup.D). We considered τ to be a change point for sample j if $\phi_{j,\tau} \neq \phi_{j,\tau+1}$. The research goal is to estimate the locations of all the change points from all samples. CNV segments can then be generated by connecting adjacent change points. Conventional single sample methods have worked on

this problem by simply applying the calling algorithm to each data sequence repeatedly (D’Aurizio *et al.*, 2016; Jiang *et al.*, 2018).

In our method, we assumed that certain change points were shared by multiple samples with population frequency p_τ ; and there were G change points in total. Let $\boldsymbol{\tau} = \{\tau_1, \dots, \tau_G\}$ denote the locations of those shared change points. For each $\tau_g (g = 1, \dots, G)$, we considered the j -th sample as a CNV carrier if $\phi_{j,\tau_g} \neq \phi_{j,\tau_g+1}$. Therefore, the goal is now to estimate all the sample shared CNV regions (i.e., $\boldsymbol{\tau}$) and then identify individual carriers. We used a two-stage method described below. Stage I uses a probit stick-breaking process to identify shared CNV regions (Section 4.2.2) and initially filters locations without CNVs (Section 4.2.4). Stage II calls CNVs individually (Section 4.2.3).

4.3.2 STAGE-I: SHARED CNV INFERENCE BY BAYESIAN PROBIT STICK-BREAKING PROCESS MODEL

We modelled the corresponding latent means and variances ϕ_i using the Bayesian model through a probit stick-breaking process (PSBP) (Chung and Dunson, 2009; Rodríguez and Dunson, 2011). Unlike other non-parametric processes (e.g., Dirichlet process), PSBP is immune to the centering problems, which guarantees the mean of latent variable (i.e., ϕ_i) is centered at a specific level, making it ideal to model ϕ_i (Cai and Bandyopadhyay, 2017). The PSBP also has a nice shrinkage property, allowing for efficiently clustering high dimensional ϕ_i to a small number of clusters (i.e., copy number states). Moreover, the PSBP mixture model can capture multimodal and heavy-tailed distribution, which relaxed the normality assumption of latent ϕ_i , providing more flexible scenarios for modelling the complex CNV data. Specifically, we assumed ϕ_i followed an unknown distribution $G \sim PSBP(\alpha G_0)$ with centering distribution G_0 and

shape measure α that reflected how far away the random distribution G is from the center G_0 . Following Rodriguez et al., G admitted a representation of the form (Rodríguez and Dunson, 2011):

$$\phi_i \sim G(.) = \sum_{l=1}^L \omega_l \delta_{\theta_l}(.) \quad (11)$$

where L represented the number of all possible copy number states (e.g., $L=5$), $\theta_l = (\mu_l, \sigma_l^2)$ are possible distinct mean and variance specific to each copy number state ($l = 1, 2, \dots, L$), $\delta_{\theta_l}(.)$ is a degenerate distribution at θ_l , and $\omega_l = \Phi(\alpha_l) \prod_{r < l} (1 - \Phi(\alpha_r))$ represented the probability of assigning θ_l to each position where $\Phi(.)$ is the probit function and $\alpha_l \sim N(\mu_\alpha, \sigma_\alpha^2)$. Following this structure, each ϕ_i was assigned to one of the $\{\theta_l\}$ based on the observed intensities across all potential carriers of the copy number state for locus i . The carriers were initially identified using the strategy described below in Section 2.4. To simultaneously implement the variable selection and clustering procedures for the purpose of CNV detection, we further reconstructed the PSBP model as follows (George and McCulloch, 1993; Cai and Bandyopadhyay, 2017):

$$\phi_i \sim \gamma_i G_{\mu=0} + (1 - \gamma_i) G(.) \quad (12)$$

where the $G_{\mu=0}$ was the underlying distribution of the normal copy number states with the mean fixed at zero (i.e., diploids). $\gamma_i \sim \text{Bernoulli}(\kappa)$ is an indicator of ϕ_i being in $G_{\mu=0}$ (i.e., normal state) or not, which incorporated variable selection of the locus across samples. Specifically, when $\gamma_i = 1$, ϕ_i followed a distribution $G_{\mu=0}$; whereas $\gamma_i = 0$ indicated a potential CNV locus following $G(.)$ defined in equation (1). Within this framework, the posterior probability of ϕ_i being $G_{\mu=0}$ or not was first calculated through inference on γ_i . While the ϕ_i given $\gamma_i = 0$ was then assigned to its most possible state

according to the posterior probabilities of belonging to various copy number states (details in Section 4.2.5).

4.3.3 STAGE-II: INDIVIDUAL CNV CALLING

After a CNV region is identified across samples, we then determine the carriers in the samples, that is, to call CNVs in each individual sample. Specifically, after obtaining the posterior mean and variance estimates specific to each copy number state (i.e., $\hat{\theta}_l$ and $\hat{\sigma}_l^2, l = 1, \dots, L$), we constructed the interval for each state as $C_l = [\hat{\theta}_l - c_{1l}\hat{\sigma}_l, \hat{\theta}_l + c_{2l}\hat{\sigma}_l]$. We considered a sample as the carrier and classified the segment into the l -th copy number state if its segmental mean fell within one specific interval C_l . Values of c_1 and c_2 could be arbitrarily chosen according to empirical evidence about the magnitude of mean shifts of each CNV state, which may vary by genotyping platforms. In practice, we will suggest plotting the genotyping signals of CNV segments that were identified under different combinations of c_1 and c_2 for visualization. True positive rate (TPR) could be calculated for each combination. The optimal choices for c_1 and c_2 would achieve the highest TPR.

4.3.4 INITIAL DATA POOLING BY GAUSSIAN MIXTURE MODEL

A major complexity of multi-sample integration is how to effectively combine the information in the presence of samples containing no CNV signals. The intensities from non-carriers will dilute the signal, which may hence significantly decrease the detection power. Li and Tseng adopted a weighted technique to downweigh the non-carriers; Sung et al. used the ordered p-values from all samples and only selected samples with small p-values for CNV inference (Li and Tseng, 2011; Song *et al.*, 2016). With the same spirit,

we considered a preliminary data filtering step so that only intensities that were most likely arising from the carriers would be selected for stage-I CNV calling.

Specifically, for the i -th position, we let $Y_{ij} = (Y_{i1}, Y_{i2}, \dots, Y_{in})^T$ denote the CNV data vector across all samples. A Gaussian mixture model with three copy number mixture states (including deletions, normal state, and duplications) was considered:

$$P(Y_{ij} = y) = \sum_{k=1}^3 \pi_k P(Y_{ij} = y | Z_{ij} = k) \quad (13)$$

Z_{ij} was the latent mixture component, π_k was the mixture proportion reflecting the probability that Y_{ij} belonged to the k -th mixture component, and $P(Y_{ij} = y | Z_{ij} = k) \sim N(\mu_k, \sigma_k^2)$ was the component distribution. All the component-specific parameters (i.e., π_k, μ_k, σ_k^2) were estimated by the expectation-maximization (EM) algorithm. Samples were then assigned to latent clusters with the largest estimates of π_k (Dempster *et al.*, 1977). Afterward, samples showing evidence of diploid (i.e., $k = 2$) will be filtered.

4.3.5 HYPERPARAMETERS CHOICES AND MCMC ALGORITHM

For the PSBP mixture model described in equations (1-2), we developed a Markov Chain Monte Carlo (MCMC) algorithm relying on a modification of the Gibbs sampler to perform the posterior inference (Ishwaran and James, 2001). Note that with the proper choices of priors and hyperpriors described below, all full conditionals are very straightforward and can be analytically derived.

We adopted the following choices for the hyperparameters (Details in Supplementary C.2). For variable selection variable $\gamma_i \sim \text{Ber}(\kappa)$, we used the Beta conjugate hyperprior for the parameter κ . For the component-specific $\theta_l = (\mu_l, \sigma_l^2)$ of $G(\cdot)$ defined in equation (15), we used the conjugate normal and inverse gamma

hyperpriors for μ_l and σ_l^2 , respectively. The posteriors of parameters can then be computed via the MCMC algorithm, for which the detailed updating steps are in Supplementary C.3. We introduced the latent variable s_i such that $s_i = l$ denoted the i -th position was assigned to the l -th component, and s_i was sampled from a multinomial distribution. To update the latent α_l and weight parameter ω_l , we adopted a data augmentation approach (Rodríguez and Dunson, 2011). We introduced a collection of conditionally independent latent variable $z_{il}(s_i) \sim N(\alpha_l, 1)$, and we defined $s_i = l$ if and only if $z_{il}(s_i) > 0$ and $z_{ir}(s_i) < 0$ for $r < l$. Therefore, we have,

$$\begin{aligned} Pr(s_i = l) &= Pr(z_{il}(s_i) > 0, z_{ir}(s_i) < 0 \text{ for } r < l) \\ &= \Phi(\alpha_l) \prod_{r < l} (1 - \Phi(\alpha_r)) = \omega_l \end{aligned} \quad (14)$$

The augmented variable $z_{il}(s_i)$ can be imputed by sampling from its full conditional distribution,

$$z_{il}(s_i) = \begin{cases} N^-(\alpha_l, 1) & l < s_i \\ N^+(\alpha_l, 1) & l = s_i \end{cases} \quad (15)$$

where N^- and N^+ denoted the negative and positive truncated normal distributions. Given $z_{il}(s_i)$, α_l can be updated from its conjugate full conditional distribution. The component-specific parameters, θ_l and σ_l^2 were also updated from their conjugate full conditionals. Finally, we updated γ_i based on the marginal likelihoods for (\mathbf{y}, \mathbf{s}) and κ from its conjugate full conditional distribution.

4.3.6 NUMERICAL SIMULATIONS

To evaluate the performance of our method, we conducted simulations under various settings. Four copy number states were simulated including del.S, del.D, dup.S, and dup.D. The CNV length varied from 10~30 markers (i.e., SNPs and exons), 30~60

markers and 60~100 markers. The CNV population frequency was set to be 20%, 50% or 100%, respectively.

First, we evaluated our method when both WES and SNP array data were available. For WES data, to generate data retaining the true noise structure and exon distribution, we conducted a spike-in design (Jiang *et al.*, 2018; Zhou *et al.*, 2018). We started with read depth data on chromosome 1 in 81 samples from the 1000 Genomes Project (Auton, Gonalo R. Abecasis, *et al.*, 2015). Exons harboring CNVs identified by EXCAVATOR2 and CODEX2 and reported in the Database of Genomic Variants (DGV) were removed (MacDonald *et al.*, 2014; D’Aurizio *et al.*, 2016; Jiang *et al.*, 2018). The read depth data of the remaining exons were treated as WES random noise background. We multiplied the background read depth by $c/2$, where c was sampled from a normal distribution with mean and variance provided in Supplementary C.6. For SNP array data, we utilized the similar strategy used in Xiao *et al.* to simulate intensities (i.e., LRR) (Xiao *et al.*, 2019), which mimicked the real data from the international HapMap consortium (D. M. Altshuler *et al.*, 2010). We then randomly selected and spiked in 50 dispersed CNV segments of varying length and frequency for every single sequence.

Using the simulated datasets, our method BMI-CNV was compared to the existing integrative method iCNV (Zhou *et al.*, 2018). The performance of methods was assessed by precision rate, recall rate, and F1 score measures (Supplementary C.6). We also evaluated the performance of our method when only WES data was available. Our method’s performance was compared against that of CODEX2, EXCAVATOR2, and iCNV in the single-platform mode (D’Aurizio *et al.*, 2016; Jiang *et al.*, 2018; Zhou *et al.*, 2018).

4.3.7 APPLICATION TO THE 1000 GENOMES PROJECT AND HAPMAP DATASETS

To further illustrate the characteristics of our method, we analyzed the same 81 individuals using SNP array and WES data from the 1000 genomes project and the international HapMap consortium. A detailed description of the experimental samples and genotyping platforms was provided in Section 2.2.5 and Section 3.2.1. Raw read counts and SNP array data were processed and normalized to generate *log2R-MED* and LRR intensities. For the WES data, we arbitrarily selected four normal samples from the 1000 genomes project as controls in the calculation of *log2R-MED* intensity (details in Supplementary C.4). The posterior inference of BMI-CNV was based on 2,000 MCMC samples with a burn-in period of 500 iterations.

4.3.8 INTEGRATIVE ANALYSIS OF TRICL CASE-CONTROL STUDY

We further applied BMI-CNV to the international lung cancer study TRICL (Amos *et al.*, 2017). We applied BMI-CNV to 1,163 samples genotyped by both OncoArray and WES data, and 829 samples that only had WES data using our integrative analysis mode and single platform analysis mode, respectively (details in Supplementary C.4 and C.5).

CNV calls were annotated by known gene regions obtained from UCSC Genome Browser (Kent *et al.*, 2002). A gene-based association test described in Section 3.3.4 was performed to investigate the influence of CNV on lung cancer susceptibility. Effects from deletions and duplication were evaluated separately, while adjusting the covariates, including smoking status (ever/never), age, gender, and top 4 principal components (Patterson *et al.*, 2006). In addition to study the overall lung cancer risk, we also performed stratification analyses by histological types of lung cancer (squamous cell lung

cancer [SQC] and lung adenocarcinoma [LUAD]). The significance of the effects from deletions and duplications were tested separately via the Wald test (i.e., $\beta_{del} = 0$, $\beta_{dup} = 0$), all nominal P-values were adjusted by Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995).

4.4 RESULTS

4.4.1 SIMULATIONS SHOWED IMPROVED PERFORMANCE OF BMI-CNV IN INTEGRATIVE ANALYSIS AND SINGLE PLATFORM ANALYSIS

We first evaluated the performance of BMI-CNV with simulated data when both SNP array and WES data were available. Various simulation settings were considered, including different CNV sizes and population frequencies. Overall, BMI-CNV outperformed iCNV in all scenarios with higher F1 scores (Figure 4.2, Table 4.1). iCNV tended to be conservative compared to our method, which maintained a high precision rate, although the recall rate was compromised. For example, when the simulated CNVs had a length of 30-60 markers and the population frequency was 20%, BMI-CNV had a precision at 0.70, a recall rate at 0.83, and an F1 score at 0.76. The corresponding values for iCNV were 0.99, 0.37, and 0.54, respectively. Moreover, at a certain value of CNV size, we observed the improved performance of BMI-CNV as CNV frequencies increased from 20% to 100%, and it achieved the highest F1 score when all the samples were carriers. The performance of the iCNV method was not sensitive to the CNV frequencies. Regarding computational speed, our method took about 280 minutes to scan a chromosome with 90,739 markers from 81 samples based on 2,000 MCMC sampling runs. The computation was performed on a regular laptop with an Intel Core i7 processor and 24.00 GB of RAM.

Next, we utilized the simulated WES data to assess the performance of BMI-CNV benchmarking against existing WES CNV detection methods. The performance of BMI-CNV was superior to other methods on detecting medium and long CNVs reflected by the largest F1 scores (Figure 4.3, Table 4.2). CODEX2 performed the best for detecting short CNVs, except for detecting low-frequency CNVs (i.e., frequency=20%). iCNV and EXCAVATOR2 tended to be conservative, as they achieved a high precision rate but with a significant sacrifice on recall rate. It was also noteworthy to mention that, when the CNV size was fixed at a certain value, the performance of BMI-CNV and CODEX2 were both improved with increased CNV frequencies, and they achieved the highest F1 scores when all the samples were carriers. Still, the performance of EXCAVATOR2 and iCNV were not subject to CNV frequencies, as they mainly scan one sample at a time. The shared information from multiple samples was not utilized in the calling algorithms. In conclusion, the BMI-CNV method, which integrated information from multiple samples, presented evidently improved performance in common CNV detection for both multi-platform integration and single-platform analyses.

4.4.2 APPLICATION TO THE 1000 GENOMES PROJECT AND HAPMAP DATA

We further applied BMI-CNV to the public datasets from the 1000 genomes project and HapMap data for evaluation. In total, we identified 37, 213 CNVs from 81 samples (Figure 4.4). Among them, 28% of the CNVs have been previously reported by DGV (MacDonald *et al.*, 2014). Most CNVs tended to be short (< 20 markers) and had a frequency of less than 50%. The total number of deletions was nearly the same as that of duplications (19, 194 vs. 18, 019). Supplementary Figure A.3 showed the summary of CNVs, which suggested no difference between deletions and duplications in the CNV

length and frequency. Moreover, by integrating the SNP array data, our method recovered 4,418 CNVs that resided in the non-coding regions that would be missed by the same method using WES data alone.

Supplementary Figure A.4 illustrated one typical common deletion region, suggesting that 27 out of 81 samples were carriers of this variant. A clear pattern of the shared deletion was observed through visual inspection of the signal intensities for this region across samples, and our proposed method presented accuracy in identifying this region. We also explored an alternative data integration strategy that only used intronic SNPs from the array data. Comparing to the main method using all SNPs, integrative calling with only intronic SNPs yielded a similar number of CNVs but a lower concordance rate with DGV (26% vs. 28%), implying the advantage of utilizing information of all SNPs on slightly improved accuracy (Figure 4.4, Supplementary Figure A.5).

4.4.3 INTEGRATIVE ANALYSIS OF TRICL CASES AND CONTROLS

With the TRICL datasets, we identified 253,183 CNVs in autosomes from 1,992 samples (Figure 4.5) in total. Overall, we detected more deletions than duplications, with an average length of deletions (in markers) larger than that of the duplications (13.46 markers vs. 8.72 markers) (Supplementary Table A.11). No significant difference in the overall proportion of deletions and duplications between cases and controls was observed (49% vs. 51% for deletions and 52% vs. 48% for duplications, respectively). Also, there was no difference in the length of deletions and duplications between cases and controls (Supplementary Table A.11).

Those identified CNVs were mapped to 3,472 genes. An association test with SQC subgroup highlighted the deletion gene *LGALS9* in 17q11.2 region (OR=4.14, 95% CI=1.65-10.38, *P*-value=0.002) and duplication genes *HSPG2* in 1p36.12 region (OR=4.79, 95% CI=1.75-13.10, *P*-value=0.002), *EIF3E* in 8q23.1 region (OR=2.19, 95% CI=1.31-3.64, *P*-value=0.003). Association results in the LUAD subgroup identified the duplication gene *YTHDC2* in 5q22.2 region (OR=2.88, 95% CI=1.62-5.12, *P*-value=0.0003), which was also identified in the overall lung cancer risk model by adjusting the histological subtypes as a covariate (Supplementary Table A.12). The intensities' plots indicated that all of those variants were valid CNV segments that showed distinct data patterns from other non-carriers and adjacent regions (Supplementary Figure A.6). Although these genes were not significant after multiple comparison adjustments, they still provided potential evidence and great insights into future studies on the roles of CNVs in lung cancer risk.

4.5 DISCUSSION

Challenges shared by existing WES methods are the lack of sensitivity for common CNVs and the incapability of studying the non-coding regions of the genome. In this chapter, we have developed a multiple sample-based method, BMI-CNV, to improve common CNV detection with WES data, allowing for the integration of available SNP array data. The simulation results demonstrated the desirable performance across different scenarios of CNV sizes and population frequencies. The improvement for calling long and high-frequency CNVs was the most substantial. We reanalyzed the WES data from the 1000 Genomes Project and SNP array data previously generated by HapMap project 3 and demonstrated the advantage of multi-platform integration over the

single-platform analysis. Finally, our application of BMI-CNV to WES and OncoArray datasets of the TRICL consortium indicated potential lung cancer associated CNVs.

This is the first report demonstrating the improved performance of CNV detection by utilizing a multi-sample and multiple-platform strategy. The advantages of our method in theory are in two aspects. First, utilizing information across samples will dramatically reduce the false positives and boost the detection power. We showed that BMI-CNV presented essential advantages over other single-sample methods in detecting common variants. The advantage has been previously shown in Song et al. (Song *et al.*, 2016), which proved that the underlying statistical power of multi-sample methods converged to one at a faster rate than single-sample methods. Second, BMI-CNV integrates available SNP data to detect CNVs in non-coding regions, allowing for full spectrum genomic variants investigation. Indeed, the important role of CNVs in non-coding regions has been revealed in numerous studies. For example, Kumaran et al. identified 1,812 breast cancer-associated CNVs mapping to non-coding regions (Kumaran *et al.*, 2018). D'Aurizio et al. and Kuilman et al. developed WES-based methods, EXCAVATOR2 and CopywriteR, which used both the targeted reads and the nonspecifically captured off-target reads (i.e., from the non-coding region) (Kuilman *et al.*, 2015; D'Aurizio *et al.*, 2016). Unfortunately, the information contained in the off-targets is too biased and incomplete. Our method utilizes the more complete SNP array data from the matched samples, which provides a more reliable and unbiased solution. iCNV uses a single hidden Markov model to jointly analyze data from all platforms. It assumes that those overlapping markers (i.e., exons and SNPs) share the same copy number and indeed use one platform to validate the calls from the other. In contrast, BMI-CNV systematically

combines data sequences from multi-platforms and allows the overlapping markers to have different copy number states.

Our method still presents limitations. First, it does not detect rare CNVs, as the power will be attenuated in the existence of a large proportion of non-carriers. Second, it will have low power to detect CNVs with similar proportions of duplications and deletions in the samples, which might be less likely to happen for germline variations. Our method may split those CNVs into several smaller deletions and duplications, as each CNV locus is equally likely to be assigned to deletion or duplication. Thirdly, our data integration strategy merely combines multi-platform datasets into one data sequence instead of using them as separate sources of input, which may lose some information especially for overlapping markers (e.g., exons and SNPs). One possible solution is to implement dimension reduction techniques on those datasets and perform CNV identification on the inferred latent factors, similar idea has been used for the integrative analysis of multi-omics data (Ma and Zhao, 2012; Meng *et al.*, 2016; Park *et al.*, 2021). Another solution is to implement the methodologies of multivariate change-point detection (Aue *et al.*, 2009; Kuncheva, 2013; Montañez *et al.*, 2015; Truong *et al.*, 2020). Those methods have been extensively used to identify changes in multivariate time series data.

TABLES AND FIGURES

Table 4.1 Summary of performance of our method on simulated data in integrative analysis. Precision rate, recall rate and F1 score are summarized for CNV calls generated by BMI-CNV and iCNV in integrative analysis. Frequency: CNV frequency.

		BMI-CNV			iCNV		
CNV length (markers)	Frequency	Precision	Recall	F1	Precision	Recall	F1
10~30	20%	0.54	0.71	0.61	0.98	0.35	0.51
	50%	0.58	0.90	0.71	0.98	0.35	0.51
	100%	0.79	0.91	0.84	0.98	0.35	0.52
30~60	20%	0.70	0.83	0.76	0.99	0.37	0.54
	50%	0.75	0.87	0.81	0.99	0.35	0.52
	100%	0.92	0.85	0.89	0.99	0.36	0.53
60~100	20%	0.75	0.88	0.81	0.99	0.40	0.57
	50%	0.87	0.89	0.88	0.99	0.39	0.56
	100%	0.88	0.89	0.89	0.99	0.40	0.57

Table 4.2 Summary of performance of our method on simulated data in WES analysis. Precision rate, recall rate and F1 score are summarized for CNV calls generated by BMI-CNV, iCNV, EXCAVATOR2 and CODEX2 in WES analysis. Frequency: CNV frequency.

		BMI-CNV			iCNV			EXCAVATOR2			CODEX2		
CNV length	Frequency	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
10~30	20%	0.44	0.77	0.56	0.99	0.40	0.57	0.90	0.21	0.34	0.37	0.99	0.54
	50%	0.46	0.85	0.59	0.98	0.39	0.56	0.87	0.21	0.34	0.49	0.98	0.65
	100%	0.90	0.64	0.74	0.99	0.37	0.54	0.87	0.25	0.39	0.97	0.63	0.76
30~60	20%	0.81	0.55	0.65	0.99	0.42	0.59	0.80	0.19	0.30	0.35	0.99	0.51
	50%	0.65	0.79	0.71	0.99	0.42	0.59	0.81	0.20	0.32	0.48	0.91	0.63
	100%	0.82	0.79	0.81	0.99	0.43	0.60	0.84	0.25	0.38	0.99	0.69	0.81
60~100	20%	0.78	0.72	0.75	0.99	0.54	0.70	0.93	0.26	0.41	0.25	0.74	0.37
	50%	0.80	0.77	0.78	0.99	0.53	0.69	0.87	0.31	0.46	0.45	0.42	0.43
	100%	0.87	0.78	0.83	0.99	0.54	0.70	0.90	0.39	0.54	0.99	0.68	0.81

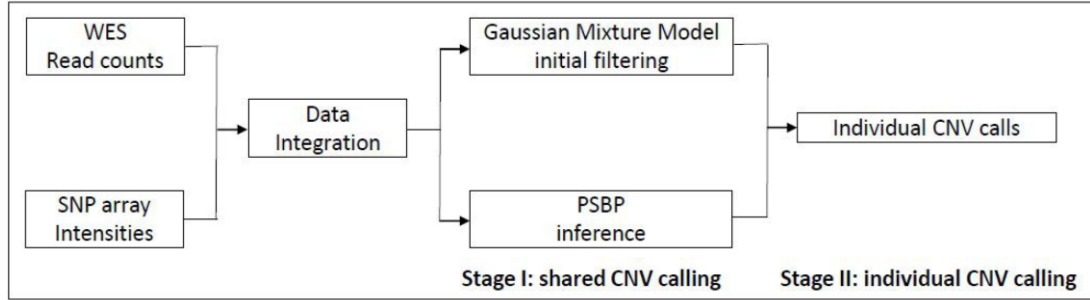


Figure 4.1 Analysis workflow of BMI-CNV. BMI-CNV requires two inputs: (1) WES raw read count data from testing and control samples that are computed by using genotyping tools such as SAMtools; (2) SNP array intensities. WES read counts are normalized to correct exon length, GC-content and mappability biases. Logarithm of normalized values between testing and pooled control samples are calculated. SNP array intensities are normalized to adjust the genomic waves. The WES and SNP array data are standardized by robust scaling approach and then integrated. For CNV calling, BMI-CNV carries out a two-stage framework to generate CNV calls. In stage I, an initial data filtering procedure is coupled with a Bayesian PSBP method to identify shared CNV regions. In stage II, an individual CNV calling procedure is performed to call CNVs in each sample.

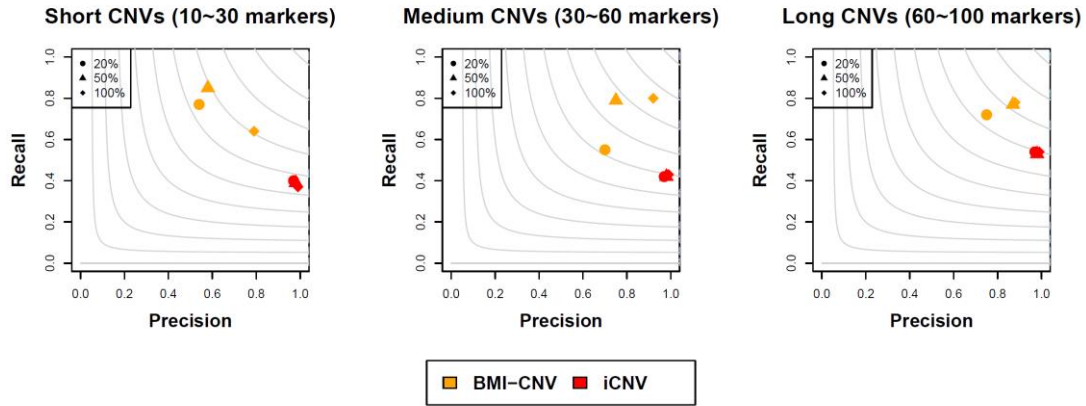


Figure 4.2 Performance assessment of BMI-CNV and iCNV on simulated data in integrative analysis. Simulated CNVs are of frequency 20%, 50% and 100% and length 10~30 markers (short), 30~60 markers (medium) and 60~100 markers (long). The grey contours are F1 scores calculated as the harmonic mean of precision and recall rates.

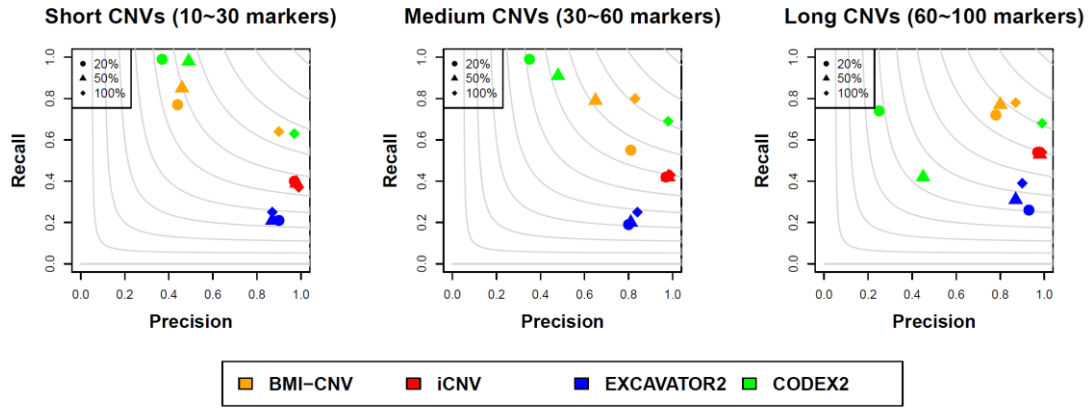


Figure 4.3 Performance assessment of BMI-CNV, iCNV, EXCAVATOR2 and CODEX2 on simulated data in WES analysis. Simulated CNVs are of frequency 20%, 50% and 100% and length 10~30 markers (short), 30~60 markers (medium) and 60~100 markers (long). The grey contours are F1 scores calculated as the harmonic mean of precision and recall rates.

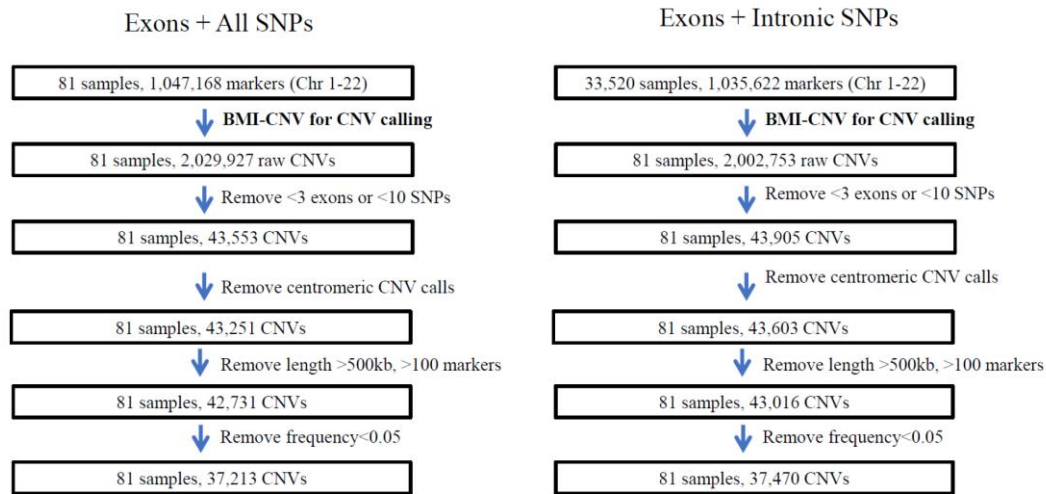


Figure 4.4 Overview of the application to the 1000 genomes project and HapMap data. The figure outlines the study design with brief description of quality control (QC) methods. Summary of key results include the sample size and number of CNVs at various stages of analysis. Left: CNV calling results using all SNPs and exons; right: CNV calling results using intronic SNPs and exons. Chr: chromosome.

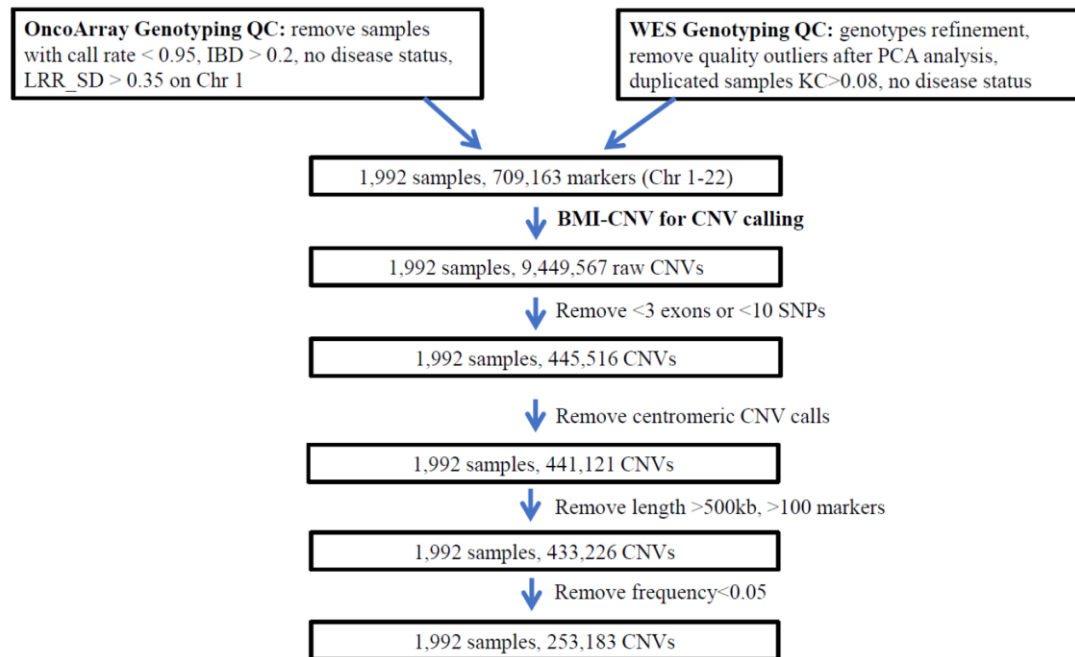


Figure 4.5 Overview of the integrative analysis of TRICL case-control study. The figure outlines the study design with brief description of quality control (QC) steps. Summary of key results include the sample size and number of CNVs at various stages of analysis. IBD: identical by descent; KC: kinship coefficient; LRR_SD: standard deviation of Log R ratio; Chr: chromosome; PCA: principal component analysis.

CHAPTER 5

CONCLUSIONS AND FUTURE PERSPECTIVES

CNVs are an important and pervasive source of genetic variation that accounts for a significant proportion of inherited susceptibility to human diseases. Understanding the mechanisms underlying the influences of CNVs will be instrumental for many basic research areas. Accurate identification of CNVs is highly demanded to provide a comprehensive view of human genetic variation. This dissertation developed three novel methods to improve the CNV detection, which can be implemented with data from various genotyping platforms: SNP array, WES, and combination of SNP array and WES. First, we developed a novel algorithm LDcnv. LDcnv employs a local search strategy that directly integrated the genomic correlation structure, allowing for efficient CNV identification with correlated data. Through extensive simulations and analyses of HapMap datasets, we showed that LDcnv outperforms existing methods, especially for detecting short CNVs.

To further explore the potential of correlation-based algorithm on detecting CNVs in WES data, we developed two different CNV detection tools. We evaluated the existing normalization approaches and find the median based normalization retained the correlation structure and achieved superior performance coupled with the LDcnv algorithm, so as referred to the CORRseq method. The performance of CORRseq was evaluated in extensive simulation studies and real data applications from the 1000 Genomes Project. CORRseq outperformed existing methods in detecting medium and

large CNVs. We also developed a Bayesian Multi-sample and multi-platform CNV detection method to achieve higher accuracy and resolution. By utilizing data sequenced by both WES and microarray, BMI-CNV addresses the limitation shared by single-platform methods due to the sparse nature of WES and microarray, while incorporating information across samples significantly improves the robustness of CNV identification against data noises. With extensive simulations, BMI-CNV outperformed existing methods for both multiple and single platform analyses.

These newly developed methods will be useful for the genetics and genomics community in studying CNVs with complex and high-throughput data. Genomic correlation structure is an important component in understanding and modelling the CNV data. In our work, we explored the correlation structure of the genetic intensity data from both SNP array and WES data. Strong evidence was found to support the existence of genomic dependence structure in the CNV data. Especially in the SNP array data, we theoretically proved that the correlation between two loci in the genetic intensity will depend on the LD coefficient computed from SNP allele frequencies. Both LDcnv and CORRseq demonstrated the promise of modelling the genomic correlation structure in obtaining more accurate and reliable CNV estimates with SNP array and WES data. These methods also fill in a gap in existing change-point detection methods to allow for modelling correlated observations. BMI-CNV provided a Bayesian framework to jointly investigate CNV signals from multiple samples and platforms, which has several essential advantages over other modelling strategies. First, the nonparametric PSBP can relax the restrictive parametric assumption and allows flexible modelling of the complex high-throughput data. Second, the Bayesian framework enables great flexibility and

possibility to incorporate prior relevant information such as the documented CNV hotspot information (Wang *et al.*, 2007). Finally, the PSBP framework can be easily extended to accommodate the complex data dependence structure by replacing the independent weights with stochastic processes (e.g., Gaussian process) without sacrificing computational tractability (Rodríguez and Dunson, 2011). It opens a door to exploit various sources of available information to improve the performance of CNV calling algorithms.

The application of our newly developed methods, CORRseq and BMI-CNV, to TRICL case-control study identified several candidate lung cancer risk related genes, which may provide new insights into the etiology of lung cancer. The discovered significantly associated deletion genes *AMBRA1*, *FCGR3A*, and *LGALS9* were previously found to be novel prognostic markers in lung cancer, loss and low expression of which were correlated with fast tumor growth and poor survival outcome (He *et al.*, 2019; Xu and Guo, 2020; Chaikovsky *et al.*, 2021). The significant amplification genes *HSPG2*, *EIF3E*, *YTHDC2*, *ZNF280A*, *PRAME*, and *AMY2B* also have been frequently described as oncogenes in diverse tumor types, including lung cancer, gastrointestinal cancers, and breast cancer (Yang *et al.*, 2015; Xu *et al.*, 2020; Liu *et al.*, 2021). Besides that, the regions 1p36, 3p14, and 19q13.3 were previously identified as a tumor suppression region, where loss of those tumor suppressors contributes to cancer development (Whang-Peng, 1989; Sobottka *et al.*, 2000; Carén *et al.*, 2007). Further large-scale studies are needed to validate these potential findings.

Next, we will extend our methods to be capable of analyzing single-cell DNA-sequencing (scDNA-seq) data. scDNA-seq data is ideal for inferring CNVs in each cell,

which make it possible to study various tumor characteristics, such as intra-tumor heterogeneity and tumor evolutionary trajectory. Tumors are typically heterogeneous, which comprise of several sub-clonal populations (Navin and Hicks, 2010; McGranahan and Swanton, 2017). Identifying and extracting those genetically distinct sub-populations is crucial for reconstructing tumor evolutionary trajectories, which may uncover the mechanisms of tumor progression. Several studies have implicated CNVs in dissecting and clustering of tumor cells (Zack *et al.*, 2013; Velazquez-Villarreal *et al.*, 2020; Zhou *et al.*, 2020) and performing evolutionary analysis (Dorri *et al.*, 2020; Kuipers *et al.*, 2020), where a key challenge comes from low and uneven sequencing coverage that may lead to false calls (Mallory *et al.*, 2020; Lähnemann *et al.*, 2020). Another direction is to perform single-cell RNA-sequencing (scRNA-seq) in parallel to scDNA-seq using the same tumor samples, our Bayesian framework can be extended to integrate multiple data types to enhance detection power. Moreover, single cell transcriptomic profiling is promising in characterizing tumor heterogeneity and trajectories, numerous methods have been developed, see Kiselev *et al.* and Saelens *et al.* for complete reviews (Kiselev *et al.*, 2019; Saelens *et al.*, 2019). However, one issue is that different data types may provide inconsistent CNV results, where CNVs identified in one data are absent in other data types. Another complexity is that although CNVs are related to gene expression alterations, the precise quantification of association between CNVs and gene expression on a cellular scale remain to be elucidated. In addition to that, our method may also be further extended to incorporate case-control status to directly identify disease-associated CNVs in a single model. We will also explore the possibility of integrating heterogeneous measurements of biological variations, such as genetic, epigenetic, and gene expression

variations, to comprehensively assess the generic architectures of diseases. Especially with the advent of the single cell multimodal omics method, where multiple measurements can be simultaneously generated on single cells (Li *et al.*, 2019; Zhu *et al.*, 2020). However, integrative computational methods have just started to emerge, more sophisticated methods are needed to effectively handle the high level of data sparseness and noisiness and information discrepancies among different data types.

REFERENCES

- A,B. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics. *Nucleic Acids Res.*, **47**.
- Altshuler,D. *et al.* (2010) The International HapMap 3 Consortium.. Integrating common and rare genetic variation in diverse human populations. *Nature* 467: 52-58. *Nature*, **467**, 52–58.
- Altshuler,D.M. *et al.* (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*.
- Amos,C.I. *et al.* (2017) The oncoarray consortium: A network for understanding the genetic architecture of common cancers. *Cancer Epidemiol. Biomarkers Prev.*
- Asadollahi,R. *et al.* (2014) The clinical significance of small copy number variants in neurodevelopmental disorders. *J. Med. Genet.*, **51**, 677–688.
- Aue,A. *et al.* (2009) Break detection in the covariance structure of multivariate time series models. *Ann. Stat.*, **37**, 4046–4087.
- Auton,A., Abecasis,Gonçalo R., *et al.* (2015) A global reference for human genetic variation. *Nature*.
- Auton,A., Abecasis,Gonçalo R., *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
- Beecham,G.W. *et al.* (2017) Clinical/Scientific Notes: The Alzheimer’s disease sequencing project: Study design and sample selection. *Neurol. Genet.*
- Benelli,M. *et al.* (2010) A very fast and accurate method for calling aberrations in array-CGH data. *Biostatistics*, **11**, 515–518.
- Benjamini,Y. and Hochberg,Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B*.
- Cai,B. and Bandyopadhyay,D. (2017) Bayesian semiparametric variable selection with applications to periodontal data. *Stat. Med.*
- Carén,H. *et al.* (2007) Genetic and epigenetic changes in the common 1p36 deletion in neuroblastoma tumours. *Br. J. Cancer*, **97**, 1416–1424.
- Chaikovsky,A.C. *et al.* (2021) The AMBRA1 E3 ligase adaptor regulates the stability of cyclin D. *Nature*, 1–5.

- Chung,Y. and Dunson,D.B. (2009) Nonparametric bayes conditional distribution modeling with variable selection. *J. Am. Stat. Assoc.*
- Colella,S. *et al.* (2007) QuantiSNP: An objective bayes hidden-markov model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.*, **35**, 2013–2025.
- Conrad,D.F. *et al.* (2010) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**, 704–712.
- Craig Venter,J. *et al.* (2001) The sequence of the human genome. *Science* (80-.).
- D’Aurizio,R. *et al.* (2016) Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2. *Nucleic Acids Res.*
- Dempster,A.P. *et al.* (1977) Maximum Likelihood from Incomplete Data Via the EM Algorithm . *J. R. Stat. Soc. Ser. B.*
- Dorri,F. *et al.* (2020) Efficient Bayesian inference of phylogenetic trees from large scale, low-depth genome-wide single-cell data.
- Fang,H. *et al.* (2014) Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Med.*
- Fellermann,K. *et al.* (2006) A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *Am. J. Hum. Genet.*, **79**, 439–448.
- Fellermann,K. *et al.* (2003) Crohn’s disease: A defensin deficiency syndrome? *Eur. J. Gastroenterol. Hepatol.*, **15**, 627–634.
- Frank,B. *et al.* (2007) Copy number variant in the candidate tumor suppressor gene MTUS1 and familial breast cancer risk. *Carcinogenesis*, **28**, 1442–1445.
- Friguet,C. *et al.* (2009) A Factor Model Approach to Multiple Testing Under Dependence. *J. Am. Stat. Assoc.*, **104**, 1406–1415.
- Fromer,M. *et al.* (2012) Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am. J. Hum. Genet.*
- George,E.I. and McCulloch,R.E. (1993) Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.*
- Geraldine A. Van der Auwera,B.D.O. (2020) Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. 368.
- Hao,N. *et al.* (2013) Multiple change-point detection via a screening and ranking algorithm. *Stat. Sin.*, **23**, 1553–1572.

- Huang,T. *et al.* (2005) Detection of DNA copy number alterations using penalized least squares regression. *Bioinformatics*, **21**, 3811–3817.
- Ishwaran,H. and James,L.F. (2001) Gibbs sampling methods for stick-breaking priors. *J. Am. Stat. Assoc.*
- J,S. *et al.* (2007) Strong association of de novo copy number mutations with autism. *Science* (80-.), **316**, 445–449.
- Jiang,Y. *et al.* (2015) CODEX: a normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Res.*, **43**, e39.
- Jiang,Y. *et al.* (2018) CODEX2: Full-spectrum copy number variation detection by high-throughput DNA sequencing. *Genome Biol.*
- Kent,W.J. *et al.* (2002) The Human Genome Browser at UCSC. *Genome Res.*
- Kim,J.H. *et al.* (2012) CNVRuler: A copy number variation-based case-control association analysis tool. *Bioinformatics*, **28**, 1790–1792.
- Kiselev,V.Y. *et al.* (2019) Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.*, **20**, 273–282.
- Koboldt,D.C. *et al.* (2012) VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.
- Koepcke,L. *et al.* (2016) Single and multiple change point detection in spike trains: Comparison of different CUSUM methods. *Front. Syst. Neurosci.*, **10**.
- Krumm,N., Sudmant,Peter H, *et al.* (2012) Copy number variation detection and genotyping from exome sequence data. *Genome Res.*, **22**, 1525–1532.
- Krumm,N., Sudmant,Peter H., *et al.* (2012) Copy number variation detection and genotyping from exome sequence data. *Genome Res.*, **22**, 1525–1532.
- Kuilman,T. *et al.* (2015) CopywriteR: DNA copy number detection from off-target sequence data. *Genome Biol.*
- Kuipers,J. *et al.* (2020) Single-cell copy number calling and event history reconstruction.
- Kumaran,M. *et al.* (2018) Breast cancer associated germline structural variants harboring small noncoding RNAs impact post-transcriptional gene regulation. *Sci. Rep.*
- Kuncheva,L.I. (2013) Change detection in streaming multivariate data using likelihood detectors. *IEEE Trans. Knowl. Data Eng.*, **25**, 1175–1180.
- Kurt,B. *et al.* (2018) A Bayesian change point model for detecting SIP-based DDoS attacks. *Digit. Signal Process. A Rev. J.*, **77**, 48–62.
- Lähnemann,D. *et al.* (2020) Eleven grand challenges in single-cell data science. *Genome Biol.*, **21**.

- Lai, T.L. (1995) Sequential Changepoint Detection in Quality Control and Dynamical Systems. *J. R. Stat. Soc. Ser. B*, **57**, 613–644.
- Lavielle, M. and Teyssière, G. (2006) Adaptive Detection of Multiple Change-Points in Asset Price Volatility. *Long Mem. Econ.*, 129–156.
- Lesueur, F. *et al.* (2008) The contribution of large genomic deletions at the CDKN2A locus to the burden of familial melanoma. *Br. J. Cancer*, **99**, 364–370.
- Li, G. *et al.* (2019) Joint profiling of DNA methylation and chromatin architecture in single cells. *Nat. Methods*, **16**, 991–993.
- Li, H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*.
- Li, J. and Tseng, G.C. (2011) An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *Ann. Appl. Stat.*
- Lieber, M.R. (2008) The mechanism of human nonhomologous DNA End joining. *J. Biol. Chem.*, **283**, 1–5.
- Liu, H. *et al.* (2021) ZNF280A promotes lung adenocarcinoma development by regulating the expression of EIF3C. *Cell Death Dis.*, **12**.
- Luo, X. *et al.* (2020) Integrating genomic correlation structure improves copy number variations detection. *Bioinformatics*.
- Lupski, J.R. (1998) Genomic disorders: Structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.*, **14**, 417–422.
- Lupski, J.R. and Stankiewicz, P. (2002) Genome architecture, rearrangements and genomic disorders. *Trends Genet.*, **18**, 74–82.
- Ma, H. and Zhao, H. (2012) Ifad: An integrative factor analysis model for drug-pathway association inference†. *Bioinformatics*.
- MacDonald, J.R. *et al.* (2014) The Database of Genomic Variants: A curated collection of structural variation in the human genome. *Nucleic Acids Res.*
- Magi, A. *et al.* (2010) A shifting level model algorithm that identifies aberrations in array-CGH data. *Biostatistics*, **11**, 265–280.
- Magi, A. *et al.* (2011) Detecting common copy number variants in high-throughput sequencing data by using JointSLM algorithm. *Nucleic Acids Res.*, **39**, e65–e65.
- Magi, A. *et al.* (2013) EXCAVATOR: Detecting copy number variants from whole-exome sequencing data. *Genome Biol.*
- Magi, A. *et al.* (2012) Read count approach for DNA copy number variants detection. *Bioinformatics*, **28**, 470–478.

- Mallory,X.F. *et al.* (2020) Methods for copy number aberration detection from single-cell DNA-sequencing data. *Genome Biol.*, **21**.
- Mathew,B. *et al.* (2018) A novel linkage-disequilibrium corrected genomic relationship matrix for SNP-heritability estimation and genomic prediction. *Heredity (Edinb)*., **120**, 356–368.
- McCarroll,S.A. *et al.* (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.*, **40**, 1166–1174.
- McGranahan,N. and Swanton,C. (2017) Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell*, **168**, 613–628.
- Meng,C. *et al.* (2016) Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief. Bioinform.*, **17**, 628–641.
- Montañez,G.D. *et al.* (2015) Inertial hidden Markov models: Modeling change in multivariate time series. *Proc. Natl. Conf. Artif. Intell.*, **3**, 1819–1825.
- Moreno-De-Luca,D. *et al.* (2010) Deletion 17q12 is a recurrent copy number variant that confers high risk of autism and schizophrenia. *Am. J. Hum. Genet.*
- Navin,N.E. and Hicks,J. (2010) Tracing the tumor lineage. *Mol. Oncol.*, **4**, 267–283.
- Niu,Y.S. and Zhang,H. (2012) The screening and ranking algorithm to detect DNA copy number variations. *Ann. Appl. Stat.*
- Olshen,A.B. *et al.* (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*.
- Orlandini,V. *et al.* (2017) SLMSuite: A suite of algorithms for segmenting genomic profiles. *BMC Bioinformatics*.
- Packer,J.S. *et al.* (2016) CLAMMS: A scalable algorithm for calling common and rare copy number variants from exome sequencing data. *Bioinformatics*, **32**, 133–135.
- Park,S. *et al.* (2021) Integrating Multidimensional Data for Clustering Analysis With Applications to Cancer Patient Data. *J. Am. Stat. Assoc.*, **116**, 14–26.
- Patterson,N. *et al.* (2006) Population structure and eigenanalysis. *PLoS Genet.*
- Peiffer,D.A. *et al.* (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.*, **16**, 1136–1148.
- Petersen,B.S. *et al.* (2017) Opportunities and challenges of whole-genome and -exome sequencing. *BMC Genet.*, **18**.
- Plagnol,V. *et al.* (2012) A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*, **28**, 2747–2754.

- Qiu,Z.W. *et al.* (2017) Genome-wide copy number variation pattern analysis and a classification signature for non-small cell lung cancer. *Genes Chromosom. Cancer*, **56**, 559–569.
- Rodgers,K. and Mcvey,M. (2016) Error-Prone Repair of DNA Double-Strand Breaks. *J. Cell. Physiol.*, **231**, 15–24.
- Rodríguez,A. and Dunson,D.B. (2011) Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Anal.*
- Rousseeuw,P.J. and Croux,C. (1993) Alternatives to the median absolute deviation. *J. Am. Stat. Assoc.*
- Saelens,W. *et al.* (2019) A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.*, **37**, 547–554.
- Sathirapongsasuti,J.F. *et al.* (2011) Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics*, **27**, 2648–2654.
- Schwarz,K. *et al.* (2003) Human severe combined immune deficiency and DNA repair. *BioEssays*, **25**, 1061–1070.
- Seiser,E.L. and Innocenti,F. (2014) Hidden markov model-based CNV detection algorithms for illumina genotyping microarrays. *Cancer Inform.*, **13**, 77–83.
- Siegmund,D. *et al.* (2011) Detecting simultaneous variant intervals in aligned sequences. *Ann. Appl. Stat.*
- Sobottka,S.B. *et al.* (2000) Frequent loss of heterozygosity at the 19p13.3 locus without LKB1/STK11 mutations in human carcinoma metastases to the brain. *J. Neurooncol.*, **49**, 187–195.
- Song,C. *et al.* (2016) The screening and ranking algorithm for change-points detection in multiple samples. *Ann. Appl. Stat.*
- Sun,W. *et al.* (2009) Integrated study of copy number states and genotype calls using high-density SNP arrays. *Nucleic Acids Res.*, **37**, 5365–5377.
- Taylor,S.J. and Letham,B. (2018) Forecasting at Scale. *Am. Stat.*, **72**, 37–45.
- Tibshirani,R. and Wang,P. (2008) Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, **9**, 18–29.
- Truong,C. *et al.* (2018) Selective review of offline change point detection methods. *arXiv*.
- Truong,C. *et al.* (2020) Selective review of offline change point detection methods. *Signal Processing*, **167**.

- Velazquez-Villarreal,E.I. *et al.* (2020) Single-cell sequencing of genomic DNA resolves sub-clonal heterogeneity in a melanoma cell line. *Commun. Biol.*, **3**.
- 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.
- Vert,J.P. and Bleakley,K. (2010) Fast detection of multiple change-points shared by many signals using group LARS. In, *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010*.
- 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.*
- Wang,R. *et al.* (2020) SCOPE: A Normalization and Copy-Number Estimation Method for Single-Cell DNA Sequencing. *Cell Syst.*, **10**, 445-452.e6.
- Wang,S. *et al.* (2019) Association of the genetic variant rs2000999 with haptoglobin and diabetic macrovascular diseases in Chinese patients with type 2 diabetes. *J. Diabetes Complications*.
- Wei,Y.-C. and Huang,G.-H. (2020) CONY: A Bayesian procedure for detecting copy number variations from sequencing read depths. *Sci. Rep.*, **10**, 10493.
- Whang-Peng,J. (1989) 3p Deletion and Small Cell Lung Carcinoma. *Mayo Clin. Proc.*, **64**, 256–260.
- Xiao,F. *et al.* (2019) An accurate and powerful method for copy number variation detection. *Bioinformatics*.
- Xiao,F. *et al.* (2015) Modified screening and ranking algorithm for copy number variation detection. *Bioinformatics*.
- Xiao,F. *et al.* (2017) modSaRa: a computationally efficient R package for CNV identification. *Bioinformatics*, **33**, 2384–2385.
- Xu,J. and Guo,Y. (2020) FCGR1A Serves as a Novel Biomarker and Correlates With Immune Infiltration in Four Cancer Types.
- Xu,Y. *et al.* (2020) The role of the cancer testis antigen PRAME in tumorigenesis and immunotherapy in human cancer. *Cell Prolif.*, **53**.
- Yang,Z. *et al.* (2015) Integrated analyses of copy number variations and gene differential expression in lung squamous-Cell carcinoma. *Biol. Res.*, **48**.
- Zack,T.I. *et al.* (2013) Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.*, **45**, 1134–1140.

- Zare,F. *et al.* (2017) An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC Bioinformatics*, **18**, 286.
- Zhang,F. *et al.* (2009) Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.*, **10**, 451–481.
- Zhang,N.R. *et al.* (2010) Detecting simultaneous changepoints in multiple sequences. *Biometrika*.
- Zhang,N.R. and Siegmund,D.O. (2007) A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, **63**, 22–32.
- Zhang,Q. *et al.* (2010) CMDS: a population-based method for identifying recurrent DNA copy number aberrations in cancer from high-resolution data. *Bioinformatics*, **26**, 464–469.
- Zheng,X. *et al.* (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*.
- Zhou,W. *et al.* (2020) Comprehensive Analysis of Copy Number Variations in Kidney Cancer by Single-Cell Exome Sequencing. *Front. Genet.*, **10**.
- Zhou,Z. *et al.* (2018) Integrative DNA copy number detection and genotyping from sequencing and array-based platforms. *Bioinformatics*.
- Zhu,C. *et al.* (2020) Single-cell multimodal omics: the power of many. *Nat. Methods*, **17**, 11–14.

APPENDIX A

SUPPLEMENTARY FIGURES AND TABLES

Table A.1 Joint genotype probabilities for two diallelic loci. The joint genotype probabilities were calculated under the Hardy-Weinberg equilibrium assumption. A and a are reference and alternate alleles in the first locus, p_A is the probability of the reference allele; B and b are the reference and alternate alleles in the second locus, p_B is the probability of the reference allele; D_{AB} is the coefficient of linkage disequilibrium between two loci.

Locus 1	Locus 2	Probabilities
AA	BB	$(p_A p_B + D_{AB})^2$
AA	Bb	$2(p_A p_B + D_{AB})(p_A(1 - p_B) - D_{AB})$
AA	bb	$(p_A(1 - p_B) - D_{AB})^2$
Aa	BB	$2(p_A(1 - p_B) - D_{AB})((1 - p_A)p_B - D_{AB})$
Aa	Bb	$2(p_A p_B + D_{AB})((1 - p_A)(1 - p_B) + D_{AB})$ $+ 2(p_A(1 - p_B) - D_{AB})((1 - p_A)p_B - D_{AB})$
Aa	bb	$2(p_A(1 - p_B) - D_{AB})((1 - p_A)(1 - p_B) + D_{AB})$
aa	BB	$((1 - p_A)p_B - D_{AB})^2$
aa	Bb	$2((1 - p_A)p_B - D_{AB})((1 - p_A)(1 - p_B) + D_{AB})$
aa	bb	$((1 - p_A)(1 - p_B) + D_{AB})^2$

Table A.2 Relationship between CNV locations and LD map. The frequency of each type of relationship between CNV location and LD was shown, the frequency was calculated based on 300 random selected CNVs from 856 high quality CNVs from the international HapMap 3 Consortium.

	Across Block	Inter-block	Hybrid	Random
Frequency	11.00%	2.00%	4.00%	83.00%

Across Block: CNV regions covering at least one LD block; Inter Block: CNV regions occurs within one LD block; Hybrid: only one breakpoint locating within LD block; Random: CNVs locating in the area with weak or no LD structure.

Table A.3 Summary of CNV calls on simulated data at $\phi = 0.1$ from all methods. True positive rates (TPRs) and false positive rates (FPRs) of LDcnv, PennCNV, CBS and SLMSuite with different CNV states and CNV sizes are shown, while the autoregressive coefficient (ϕ) was fixed at $\phi = 0.1$ which was corresponding to Pearson's correlation coefficient at 0.1. Del.d: deletion of double copies; Del.s: deletion of single copy; Dup.s: duplication of single copy; Dup.d: duplication of double copies.

CNV State	Method	CNV length (markers)					
		10~50		50~100		100~200	
		TPR	FPR	TPR	FPR	TPR	FPR
Del.d	LDcnv_eCN	1.00	0.00	0.94	<0.01	0.93	<0.01
	LDcnv_LRR	0.99	<0.01	0.99	<0.01	0.99	<0.01
	PennCNV	1.00	<0.01	1.00	0.00	1.00	<0.01
	SLMSuite	0.99	<0.01	1.00	<0.01	0.99	<0.01
	CBS	1.00	0.01	1.00	0.02	1.00	0.02
Del.s	LDcnv_eCN	0.99	0.01	0.95	0.01	0.96	0.01
	LDcnv_LRR	0.99	<0.01	0.99	<0.01	0.97	<0.01
	PennCNV	0.98	0.02	0.96	0.03	0.90	0.08
	SLMSuite	0.99	<0.01	0.99	<0.01	0.97	<0.01
	CBS	0.99	0.02	0.99	0.01	0.99	0.01
Dup.s	LDcnv_eCN	0.98	0.01	0.93	0.01	0.94	0.01
	LDcnv_LRR	0.90	0.01	0.95	0.02	0.96	0.03
	PennCNV	0.91	0.07	0.92	0.07	0.88	0.1
	SLMSuite	0.79	0.02	0.98	0.02	0.99	0.01
	CBS	0.93	0.05	0.94	0.05	0.93	0.06
Dup.d	LDcnv_eCN	1.00	<0.01	1.00	<0.01	0.98	<0.01
	LDcnv_LRR	1.00	<0.01	1.00	0.00	1.00	0.00
	PennCNV	1.00	<0.01	0.99	0.01	0.99	<0.01
	SLMSuite	1.00	0.00	1.00	0.00	1.00	0.00
	CBS	1.00	<0.01	1.00	<0.01	1.00	>0.01

Table A.4 Summary of CNV calls on simulated data at $\phi = 0.5$ from all methods. True positive rates (TPRs) and false positive rates (FPRs) of LDcnv, PennCNV, CBS and SLMSuite with different CNV states and CNV sizes, the autoregressive coefficient (ϕ) was fixed at $\phi = 0.5$ which was corresponding to Pearson's correlation coefficient at 0.5. Del.d: deletion of double copies; Del.s: deletion of single copy; Dup.s: duplication of single copy; Dup.d: duplication of double copies.

CNV State	Method	CNV length (markers)					
		10~50		50~100		100~200	
		TPR	FPR	TPR	FPR	TPR	FPR
Del.d	LDcnv_eCN	0.99	0.01	0.96	0.01	0.99	0.03
	LDcnv_LRR	0.95	0.02	0.97	0.03	0.94	0.02
	PennCNV	0.99	0.01	0.99	0.01	1.00	<0.01
	SLMSuite	0.99	0.03	0.99	0.02	0.99	0.01
	CBS	1.00	0.23	1.00	0.44	1.00	0.64
Del.s	LDcnv_eCN	0.96	0.06	0.95	0.08	0.96	0.09
	LDcnv_LRR	0.94	0.08	0.98	0.10	0.96	0.10
	PennCNV	0.89	0.05	0.88	0.09	0.83	0.14
	SLMSuite	0.88	0.05	0.98	0.02	0.99	0.02
	CBS	0.94	0.26	0.94	0.46	0.95	0.62
Dup.s	LDcnv_eCN	0.88	0.12	0.89	0.09	0.91	0.11
	LDcnv_LRR	0.74	0.24	0.70	0.30	0.71	0.30
	PennCNV	0.84	0.11	0.86	0.13	0.83	0.18
	SLMSuite	0.56	0.08	0.91	0.09	0.92	0.08
	CBS	0.69	0.32	0.80	0.57	0.79	0.76
Dup.d	LDcnv_eCN	0.99	0.01	0.96	0.01	0.99	0.03
	LDcnv_LRR	0.99	0.01	0.98	0.02	0.97	0.03
	PennCNV	0.99	0.01	0.99	0.01	1.00	<0.01
	SLMSuite	0.99	0.02	0.99	<0.01	1.00	<0.01
	CBS	1.00	0.23	1.00	0.44	1.00	0.64

Table A.5 Overall assessment of CNV calling on the HapMap project dataset. Performance assessment of CNV calls from the HapMap Project 3 in the 180 HapMap samples by LDcnv, PennCNV, CBS and SLMSuite on reports from (a) HapMap3 (b) Conrad et al. (c) McCarroll (MCC) et al. studies. The recall rate was defined as the ratio of identified true positives over the total number of “true CNVs”. The F1 score was calculated as harmonic mean of precision rate and recall rate. TP: True positives among the detected CNVs.

	HapMap3				Conrad				MCC			
	TP	Precision	Recall	F1	TP	Precision	Recall	F1	TP	Precision	Recall	F1
LDcnv	7016	35.76%	35.19%	35.48	9999	47.58%	8.23%	14.04	4560	46.96%	38.12%	42.08
PennCNV	3760	53.23%	18.81%	27.85	4880	64.56%	4.01%	7.56	2640	66.48%	22.07%	33.14
CBS	3965	55.17%	19.88%	29.23	5025	65.08%	4.13%	7.80	2532	64.72%	21.16%	32.00
SLMSuite	4942	53.24%	24.78%	33.00	6213	62.44%	5.00%	9.00	3136	62.00%	26.21%	36.63

Table A.6 Assessment of calling performance in short CNVs on the HapMap project dataset. Performance assessment on detecting short CNVs (<10 markers) from the HapMap Project 3 in the 180 HapMap samples by LDcnv, PennCNV, CBS and SLMSuite on reports from (a) HapMap3 (b) Conrad et al. (c) McCarroll (MCC) et al. studies. The recall rate was defined as the ratio of identified true positives over the total number of “true CNVs”. The F1 score was calculated as harmonic mean of precision rate and recall rate. TP: True positives among the detected CNVs.

	HapMap3				Conrad				MCC			
	TP	Precision	Recall	F1	TP	Precision	Recall	F1	TP	Precision	Recall	F1
LDcnv	1849	8.00%	18.48%	11.17	4167	15.33%	4.23%	6.63	1390	11.86%	26.34%	16.36
PennCNV	177	2.65%	1.76%	2.12	698	8.78%	0.70%	1.31	206	5.33%	3.90%	4.51
CBS	1249	8.49%	12.48%	10.11	2879	16.73%	2.90%	4.94	834	10.50%	15.80%	12.61
SLMSuite	698	6.94%	6.97%	6.95	1398	12.07%	1.42%	2.54	540	9.65%	10.23%	9.93

Table A.7 Demographic characteristics of the participants studies after quality control filters.

		Lung cancer cases		Controls	
		Number	%	Number	%
Passed QC		1,033	55	875	45
Age					
	<=50 years	292	28	277	32
	>50 years	741	72	598	68
Sex					
	Male	607	59	511	58
	Female	426	41	364	42
Smoking status					
	Never	124	12	304	35
	Ever	6	1	3	1
	Former	413	40	373	43
	Current	484	47	192	21
Histology					
	Adenocarcinoma	548	53	875	-
	Squamous cell carcinoma	332	32	875	-
	Small cell carcinoma	33	3	875	-

Table A.8 Summary of joint CORRseq and EXCAVATOR2 calling results on TRICL data.

Status	N sample	N del	N dup	Del mean size (exons)	Del mean size (kb)	Dup mean size (exons)	Dup mean size (kb)
Cases	1,033	60,997	28,841	16.36	176.61	23.26	235.06
Controls	875	52,096	24,128	16.58	179.13	23.17	236.27
Total	1,908	113,093	52,969	16.46	177.77	23.22	235.62

N sample, number of samples; N del, number of deletions; N dup, number of duplications; Del, deletions, dup, duplications.

Table A.9 Top significantly associated CNVs with lung cancer risk (P-value<0.01)

Gene	Chr	CNV	Case:Cont	OR (95% CI)	P-value
ZNF280A	22q11.22	Duplication	700:529	1.40(1.14, 1.70)	0.001
LL22NC03-63E9.3	22q11.22	Duplication	708:536	1.38(1.13, 1.69)	0.001
PRAME	22q11.22	Duplication	705:535	1.37(1.12, 1.68)	0.001
UBXN10	1p36.12	Duplication	108:62	1.58(1.12, 2.23)	0.009
AMY2B	1p21.1	Duplication	71:31	1.80(1.14, 2.82)	0.01
LOC100130964	8p11.22	Deletion	107:59	1.76(1.23, 2.50)	0.001
FCGR3A	1q23.3	Deletion	76:31	2.00(1.28, 3.13)	0.002
LRRC46	17q21.32	Deletion	37:15	2.56(1.33, 4.93)	0.004
MRPL10	17q21.32	Deletion	53:25	2.01(1.20, 3.37)	0.007
AMBRA1	11p11.2	Deletion	20:6	3.71(1.40, 9.83)	0.008
NRAP	10q25.3	Deletion	173:108	1.45(1.10, 1.91)	0.009
NPEPPS	17q21.32	Deletion	227:152	1.37(1.08, 1.75)	0.01

Chr, chromosome; Case, number of CNVs in cases; Cont, number of CNVs in controls; OR, odds ratio; 95% CI, 95% confidence interval.

Table A.10 Top significantly associated CNVRs with lung cancer risk (P-value<0.01)

CNVR	Chr	CNV	Case:Cont	OR (95% CI)	P-value	Genes
CNVR_4385	22q11.22	Duplication	713:542	1.36(1.11, 1.66)	0.002	IGLL5,PRAME,GGTLC2,ZNF280B, ZNF280A,LL22NC0363E9.3, POM121L1P,MIR5571,MIR650,BMS1P20
CNVR_3148	13q14.11	Duplication	48:19	2.09(1.19, 3.67)	0.01	DGKH,AKAP11
CNVR_2882	12p13.2	Duplication	21:6	3.58(1.29, 9.88)	0.01	PRB2,PRB1
CNVR_5	1p36.33	Deletion	110:63	1.59(1.13, 2.24)	0.008	ATAD3A,SSU72,TMEM240
CNVR_3042	12q23.3	Deletion	25:8	3.04(1.30, 7.09)	0.01	WSCD2
CNVR_3829	17q21.32	Deletion	553:423	1.28(1.06, 1.55)	0.01	EFCAB13,KPNB1,NPEPPS,THCAT158, MRPL45P2,TBKBP1,ITGB3
CNVR_2509	10q25.3	Deletion	172:108	1.43(1.08, 1.88)	0.01	HABP2,NRAP
CNVR_4010	19p13.3	Deletion	119:67	1.53(1.09, 2.13)	0.01	WASH5P,MIR1302-2, MIR1302-9,MIR1302-10,MIR1302-11,FAM138A,FAM138F,FAM138C, LINC01002,OR4F17
CNVR_859	3p14.3	Deletion	31:15	2.35(1.20, 4.62)	0.01	WNT5A,ERC2,LINC02017,ERC2-IT1,MIR3938,CACNA2D3

Chr, chromosome; Case, number of CNVs in cases; Cont, number of CNVs in controls; OR, odds ratio; 95% CI, 95% confidence interval; genes, genes mapped to CNVRs.

Table A.11 Summary of BMI-CNV calling results on TRICL data.

Methods		N sample	N del	N dup	Del Mean size (markers)	Del Mean size (kb)	Dup Mean size (markers)	Dup Mean size (kb)
BMI-CNV	Cases	1,075	66,705	62,440	12.91	32.30	8.38	33.21
	Control	917	68,560	55,478	14.00	30.64	9.09	31.72
	Total	1,992	135,265	117,918	13.46	31.46	8.72	32.51

N sample, number of samples; N del, number of deletions; N dup, number of duplications; Del, deletions, dup, duplications.

Table A.12 Top significantly associated CNVs with lung cancer risk (P-value<0.005).

Gene	Stratum	CNV	Chr	Case:Cont	OR (95% CI)	P-value	P.adj
YTHDC2	LUAD	Duplication	5q22.2	37:22	2.88(1.62,5.12)	0.0003	0.24
MROH1	LUAD	Duplication	8q24.3	26:12	3.58(1.70,7.55)	0.001	0.24
NPEPPSP1	LUAD	Deletion	17q12	308:401	1.55(1.21,1.98)	0.0005	0.07
LOC101929950	LUAD	Deletion	17q12	308:401	1.55(1.21,1.98)	0.0005	0.07
HSPG2	SQC	Duplication	1p36.12	12:10	4.79(1.75,13.10)	0.002	0.99
EIF3E	SQC	Duplication	8q23.1	33:54	2.19(1.31,3.64)	0.003	0.99
ACAD11	SQC	Duplication	3q22.1	20:27	2.69(1.39,5.20)	0.003	0.99
HMGA2	SQC	Duplication	12q14.3	15:11	3.50(1.44,8.53)	0.005	0.99
LGALS9	SQC	Deletion	17q11.2	11:17	4.14(1.65,10.38)	0.002	0.21
COG3	SQC	Deletion	13q14.13	19:30	2.74(1.39,5.38)	0.004	0.21
TESK2	SQC	Deletion	1p34.1	24:34	2.45(1.34,4.48)	0.004	0.21
FUBP1	SQC	Deletion	1p31.1	6:5	8.65(1.91,39.18)	0.005	0.21
TBC1D23	SQC	Deletion	3q12.1	12:4	5.64(1.66,19.15)	0.005	0.21
YTHDC2	LC	Duplication	5q22.2	54:22	2.43(1.42,4.18)	0.001	0.28
MROH1	LC	Duplication	8q24.3	35:12	2.89(1.41,5.93)	0.004	0.56
HSPG2	LC	Duplication	1p36.12	23:10	3.14(1.43,6.88)	0.004	0.56
SCARF2	LC	Duplication	22q11.21	56:21	2.19(1.27,3.77)	0.005	0.56
FAM230J	LC	Duplication	22q11.21	56:21	2.19(1.27,3.77)	0.005	0.56
RIMBP3	LC	Duplication	22q11.21	56:21	2.19(1.27,3.77)	0.005	0.56
CRIPT	LC	Duplication	2p21	26:8	3.28(1.43,7.49)	0.005	0.56

P.adj is the adjusted P-values using Benjamini-Hochberg procedure; LUAD, lung adenocarcinoma; SQC, squamous cell lung cancer; LC, lung cancer; Chr, chromosome; Case, number of CNVs in cases; Cont, number of CNVs in controls; OR, odds ratio; 95% CI, 95% confidence interval.

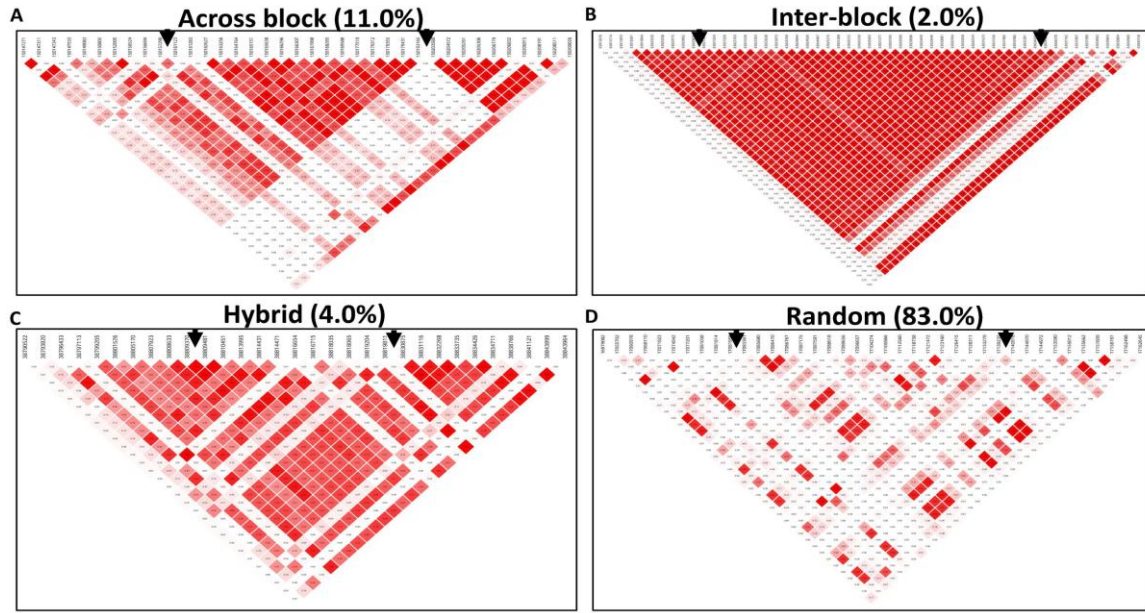


Figure A.1 Four classifications of the CNV locations in the LD genome map. The graphs summarized the frequency of CNV types with the existing high-quality CNVs from the HapMap phase 3 project. (a) Across block: CNVs spanning at least one LD blocks, (b) Inter-block: CNVs locating within a LD block, (c) Hybrid: only one breakpoint locating within LD block, and (d) Random: CNVs locating in the area with weak or no LD structure. The black arrows in each plot represent the start and end points of the CNV

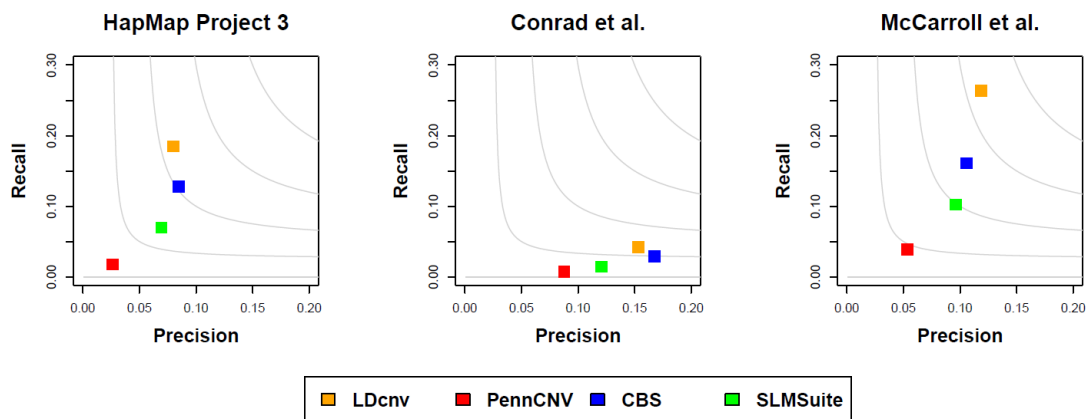


Figure A.2 Assessment of CNV calling performance in short CNVs on the HapMap project datasets. Performance assessment on detecting short CNVs (<10 markers) from the HapMap Project 3 in the 180 HapMap samples by LDcnv, PennCNV, CBS and SLMSuite on reports from (a) HapMap3 (b) Conrad et al. (c) McCarroll (MCC) et al. studies. The grey contours are F1 scores calculated as the harmonic mean of precision rate and recall rate.

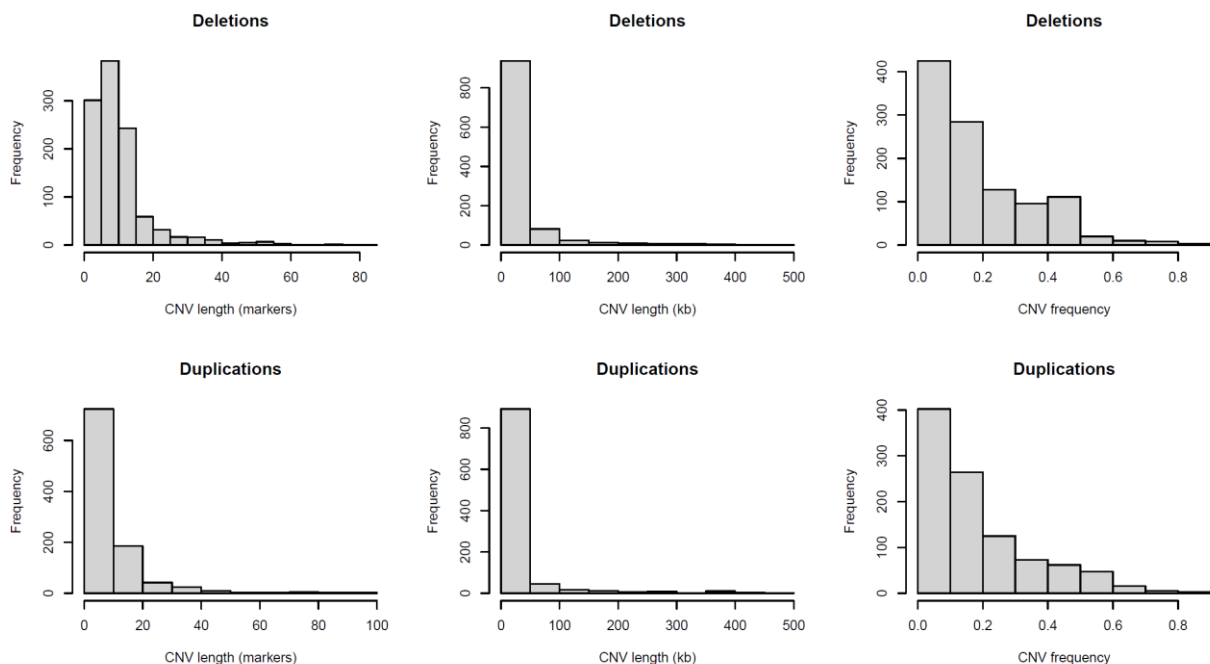


Figure A.3 Summary of length and frequency of BMI-CNV calling results on 1000 genome and HapMap project data. Genomic length (in markers and kb) and population frequency of CNVs are compared between deletions and duplications. CNVs tend to be short in size with frequency less than 50%, whereas there is no difference between deletions and duplications.

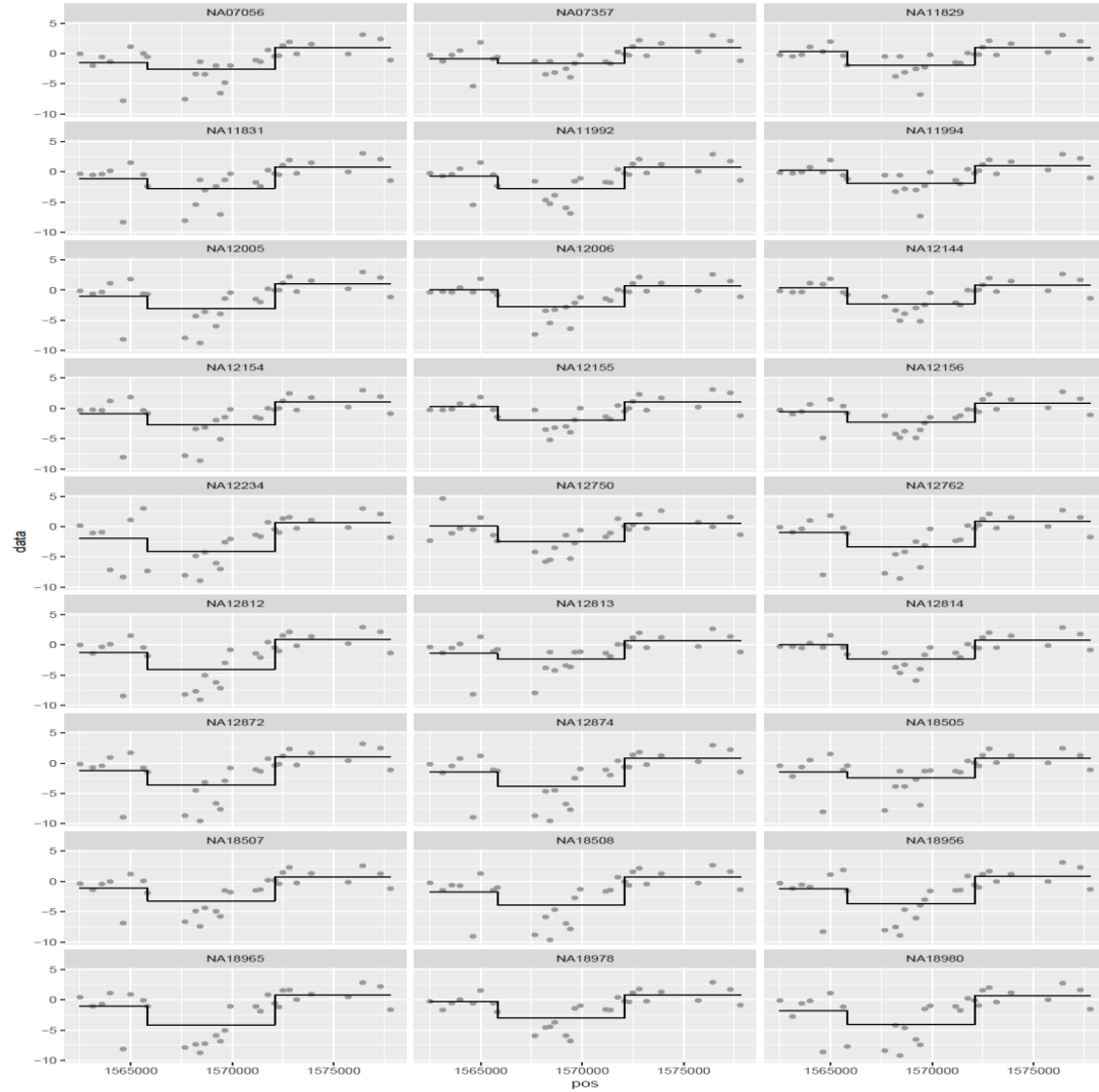


Figure A.4 Case illustration of one common deletion region of chromosome 3. The plot shows a common deletion region identified by BMI-CNV at position 52,404,727-52,425,389 of chromosome 3, this variant is shared across 27 samples. Each plot is a sample, x-axis is the genomic position and y-axis is the signal intensity (i.e. LRR or $\log_2 R-MED$). The mean signal intensities are shown by bold black lines. A clear shared deletion pattern is observed.

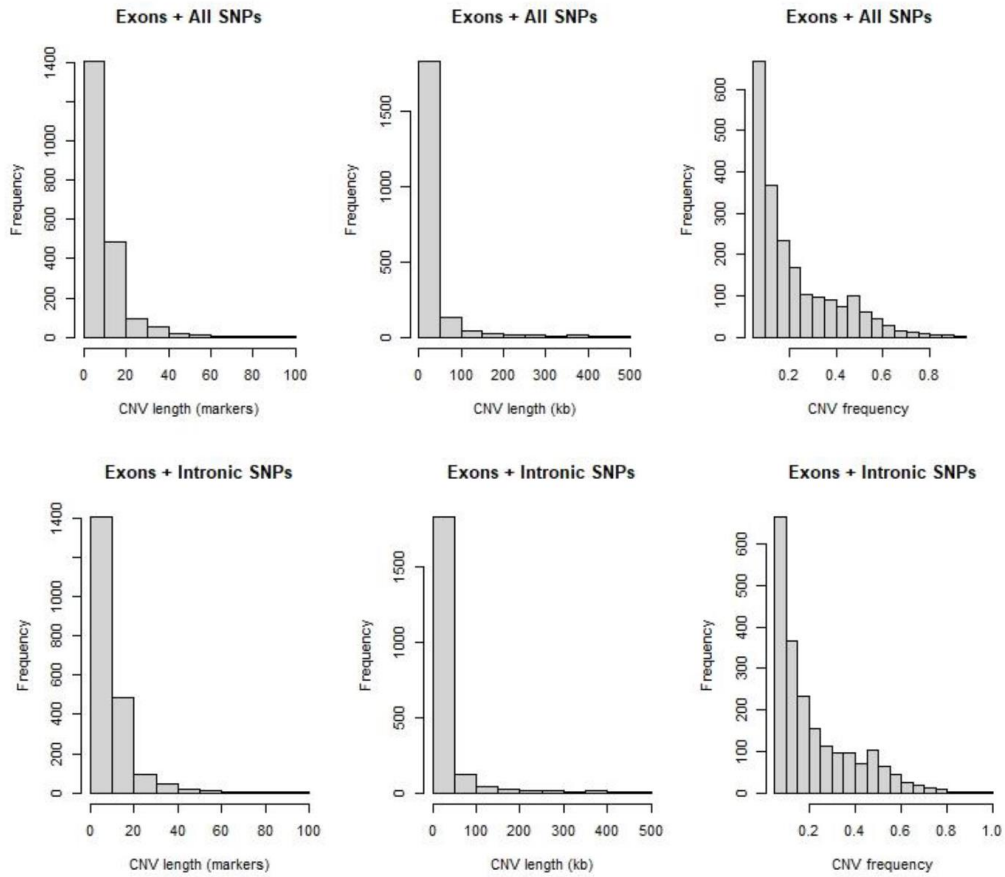


Figure A.5 Comparison of length and frequency of BMI-CNV calling results on 1000 genome and HapMap project data under two different data integration strategies. Genomic length (in markers and kb) and population frequency of CNV calling results under two data integration strategies are compared. Upper: data integration using exons and all SNPs; Lower: data integration using exons and intronic SNPs.

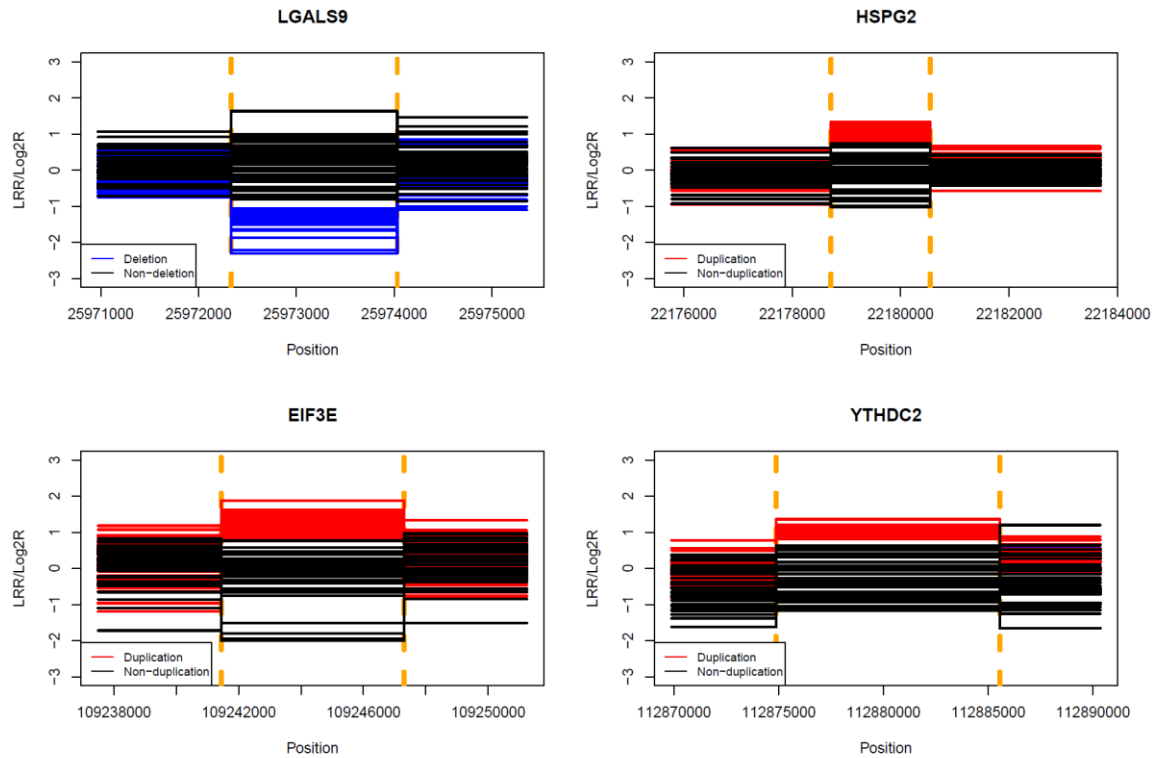


Figure A.6 Data intensity of top lung cancer related CNVs. The figure shows the data intensities of four lung cancer related CNV genes: (1) LGALS9; (2) HSPG2; (3) EIF3E; (4) YTHDC2. The mean signal intensity for each sample is shown by step bold line. For each gene, we included all carriers and other 50 randomly selected non-carriers. Vertical dashed lines depict regions identified by BMI-CNV, x-axis is the genomic position and y-axis is the signal intensity (i.e. LRR or $\log_2 R-MED$). All variants are valid CNV segments which show distinct data patterns from other non-carriers and adjacent regions.

APPENDIX B

INTEGRATING GENOMIC CORRELATION STRUCTURE IMPROVES COPY NUMBER VARIATIONS DETECTION

B.1 FALSE DISCOVERY RATE APPROACH

As discussed in (Hao *et al.*, 2013), identifying multiple CNVs or change-points is a natural multiple comparison problem. We adopted the false discovery rate (FDR) approach to adjusting the p-values in our method, as proposed by (Hao *et al.*, 2013). We used the similar strategy as used in the screening and ranking algorithm (Niu and Zhang, 2012), the test statistic $D(x)$ was calculated, naturally the larger values of which tended to support the existence of change-points. Then all local maximizers of $D(x, b)$ or, equivalently, local minimizers of p-values were selected. The locations of change-points were determined by setting a thresholding rule $|D(\hat{\tau}, w)| > \gamma$ or $p(\hat{\tau}) < p^*$. As stated in (Hao *et al.*, 2013), when the null distribution, F_0 , of the local minimizers of p-values is known, adjusted p-values could be calculated by $F_0^{-1}(p(\hat{\tau}))$. Conventional FDR procedure, such as Benjamini-Hochberg (Benjamini and Hochberg, 1995), could then be directly applied to determine the thresholding values (i.e., p^*), which guaranteed the FDR was controlled at a designated level. In the scenario when F_0 is unknown, it could be empirically estimated by generating a very long sequence of $D(x) \sim N(0, \tilde{a}\tilde{\Sigma}\tilde{a}^T)$ random variables and collected the resulting local minimizers as \widehat{F}_0 . In practice, the covariance

matrix $\tilde{\Sigma}$ was unknown but can be accurately estimated due to the sparsity of change points.

B.2 DETAILS OF THE DERIVATION OF $cov(X_A + Y_A, X_B + Y_B)$

For $cov(X_A + Y_A, X_B + Y_B)$, we obtained

$$cov(X_A + Y_A, X_B + Y_B) = cov(X_A, X_B) + cov(X_A, Y_B) + cov(Y_A, X_B) + cov(Y_A, Y_B). \quad (1)$$

To calculate each term in (1), we define the corresponding bivariate normal distributions conditional on the genotype as follow (using $f_{X_A, X_B}(x_A, x_B)$ as an example):

$$f_{X_A, X_B}(x_A, x_B) = \sum_{k=1}^4 f_{X_A, X_B}(x_A, x_B | G) P(G = G_k) \quad (2)$$

Where $G = [AABB, AABb, AaBB, AaBb]^T$ which have all the genotypes that contain the alleles A and B (Table B.1).

Table B.1 All possible Bivariate distributions and assigned Genotypes. The bivariate distributions f_s involve in calculating $cov(X_A + X_a, X_B + X_b) = cov(X_A, X_B) + cov(X_A, X_b) + cov(X_a, X_B) + cov(X_a, X_b)$ are listed, joint genotypes that are associated with each bivariate distribution are also shown under the negligibility assumption.

Covariance	Bivariate Distributions	Genotypes
$cov(X_A, X_B)$	$f_{X_A, X_B}(x_A, x_B)$	(AA, BB), (AA, Bb), (Aa, BB), (Aa, Bb)
$cov(X_A, X_b)$	$f_{X_A, X_b}(x_A, x_b)$	(AA, Bb), (AA, bb), (Aa, Bb), (Aa, bb)
$cov(X_a, X_B)$	$f_{X_a, X_B}(x_a, x_B)$	(Aa, BB), (Aa, Bb), (aa, BB), (aa, Bb)
$cov(X_a, X_b)$	$f_{X_a, X_b}(x_a, x_b)$	(Aa, Bb), (Aa, bb), (aa, Bb), (aa, bb)

So

$$\begin{aligned}
E(X_A X_B) &= E(X_A|AA)E(X_B|BB)[P(AABB) - p_A^2 p_B^2] \\
&\quad + E(X_A|AA)E(X_B|Bb)[P(AABb) - 2p_A^2 p_B(1 - p_B)] \\
&\quad + E(X_A|Aa)E(X_B|BB)[P(AaBB) - 2p_A(1 - p_A)p_B^2] \\
&\quad + E(X_A|Aa)E(X_B|Bb)[P(AaBb) - 4p_A(1 - p_A)p_B(1 - p_B)]
\end{aligned}$$

Besides, $E(X_A) = E(X_A|AA)P(AA) + E(X_A|Aa)P(Aa)$ where $P(AA) = p_A^2$ and $P(Aa) = 2p_A(1 - p_A)$ assuming HWE. Similarly, $E(X_B) = E(X_B|BB)P(BB) + E(X_B|Bb)P(Bb)$ where $P(BB) = p_B^2$ and $P(Bb) = 2p_B(1 - p_B)$. Therefore,

$$\begin{aligned}
cov(X_A, X_B) &= E(X_A X_B) - E(X_A)E(X_B) \\
&= E(X_A|AA)E(X_B|BB)[P(AABB) - p_A^2 p_B^2] \\
&\quad + E(X_A|AA)E(X_B|Bb)[P(AABb) - 2p_A^2 p_B(1 - p_B)] \\
&\quad + E(X_A|Aa)E(X_B|BB)[P(AaBB) - 2p_A(1 - p_A)p_B^2] \\
&\quad + E(X_A|Aa)E(X_B|Bb)[P(AaBb) - 4p_A(1 - p_A)p_B(1 - p_B)] \\
&= \sum_{k=1}^4 E(X_A|G_k)E(X_B|G_k)[P(G = G_k) - q(G = G_k)]
\end{aligned}$$

As $cov(X_A, X_B) = E(X_A X_B) - E(X_A)E(X_B)$, the expected values of the normalized signal intensities X_A, X_B and the expected values of their product need to be derived. The expression of all the other genotype frequencies $P(G = G_k)$ can be found in Supplementary Table A.1.

B.3 DETAILS OF THE DERIVATION OF $\sqrt{var(X_A + Y_A)var(X_B + Y_B)}$

According to the normalization procedure of the Illumina platform (https://dnatech.genomecenter.ucdavis.edu/wpcontent/uploads/2013/06/illumina_gt_normalization), the normalized intensities of the two alleles X and Y can be decomposed into

the summation of two independent raw intensities u and v . Therefore, we can express the raw intensities of the two loci as follows:

$$X_A = au + bv, Y_A = cu + dv \quad (3)$$

$$X_B = a'u' + b'v', Y_B = c'u' + d'v' \quad (4)$$

where $a, a', b, b', c, c', d, d'$ are normalizing constants. Accordingly, the denominator part of ρ_{AB} could be calculated as:

$$\sqrt{\text{var}(X_A + Y_A)\text{var}(X_B + Y_B)} = \sqrt{\pi_1 \text{var}(X_A) + \pi_2 \text{var}(Y_A)} \sqrt{\pi_3 \text{var}(X_B) + \pi_4 \text{var}(Y_B)} \quad (5)$$

where $\pi_1 = 1 + ac$, $\pi_2 = 1 + bd$, $\pi_3 = 1 + a'c'$, $\pi_4 = 1 + b'd'$. It turns out that the formula above does not depend on the correlation of the raw signal intensities of the two loci.

B.4 DETAILS OF THE INTEGRATION OF LRR AND BAF

We consider a diallelic locus A on a chromosome with two alleles, A_1 and A_2 , let $\mathbf{G}^* = (G_1^*, \dots, G_n^*)^T$ denote the genotype of a locus with eight possible genotypes and $\mathbf{S}^* = (S_1^*, \dots, S_n^*)^T$ be the underlying copy number 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). To integrate BAF information, we first introduce Lesser Allele Frequency (LAF) which it is less sparse compared to BAF:

$$LAF = \begin{cases} BAF & \forall BAF \leq 0.5 \\ 1 - BAF & \forall BAF > 0.5 \end{cases} \quad (6)$$

We then model LRRs and LAFs for each of different copy number scenarios as bivariate Gaussian distributions, with the empirical estimates of mean μ and variance σ^2

were embedded in the software `cnvPartition` (https://www.illumina.com/documents/products/technotes/technote_cnv_algorithms.pdf).

The likelihood of observing LRR and BAF for a given locus under each model is calculated via bivariate normal density, except for double deletion, the likelihood is calculated by using LRR alone. These genotype likelihoods for i -th locus are summarized by five composite copy number likelihoods: $L_{s_k,i} = \sum_{G_i \in s_k} L_{G_i}(k = 1, 2, \dots, 5)$. As a result, for the i -th locus, the preliminary copy number estimate eCN is defined as (Xiao, et al., 2019).

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

APPENDIX C

BMI-CNV: A BAYESIAN FRAMEWORK FOR MULTIPLE GENOTYPING PLATFORMS DETECTION OF COPY NUMBER VARIANTS

C.1 DATA NORMALIZATION AND STANDARDIZATION

We implemented a three-step median normalization procedure to mitigate the effect of three main observed sources of bias: exon length, GC-content and mappability (D’Aurizio *et al.*, 2016). Let $RC_{k'j}^w$ denote the raw read depth for exon k' in sample j , each $RC_{k'j}^w$ was then normalized according to:

$$\widehat{RC}_{k'j}^w = RC_{k'j}^w \times \frac{m}{m_x} \quad (8)$$

where m is the overall median, m_x is median of all the exons with the same values of exon length, mappability and GC-content. We applied this normalization procedure to both test and control samples. All pre-specified control samples were pooled together by averaging reads on each exon across all samples to form the common reference baseline. Finally, we calculated the \log_2 -ratio of normalized read counts between test samples and the reference baseline ($\log 2R-MED$).

To bring SNP array LRR and WES derived $\log 2R-MED$ to the same measuring scale, we standardized each via a robust scaling approach to produce \hat{y}_{kj}^s and $\hat{y}_{k'j}^w$. Specifically, let y_{kj}^s denote the LRR data corresponding to k -th SNP marker in sample j ; let $y_{k'j}^w$ denote the normalized $\log 2R-MED$ for exon k' in sample j (Rousseeuw and Croux, 1993), then

$$\hat{y}_{k'j}^w = \frac{y_{k'j}^w - \text{median}(y_{k'j}^w)}{\text{interquartile}(y_{k'j}^w)} \text{ and } \hat{y}_{kj}^s = \frac{y_{kj}^s - \text{median}(y_{kj}^s)}{\text{interquartile}(y_{kj}^s)} \quad (9)$$

Where $\text{interquartile}(\cdot)$ equaled the difference between 75th and 25th percentiles.

C.2 FULL MODEL SPECIFICATION

We assumed a normal linear regression model, the ϕ_i was further modelled by a variable selection prior, where $G(\cdot)$ is the PSBP.

$$Y_{ij} \sim N(Y_{ij}|\phi_i) \quad (10)$$

$$\phi_i \sim \gamma_i \delta_0 + (1 - \gamma_i) G(\cdot); \delta_0 = (\mu_0, \tau_0) \quad (11)$$

$$G(\cdot) = \sum_{l=1}^L \omega_l \delta_{\theta_l}(\cdot); \theta_l = (\mu_l, \tau_l) \quad (12)$$

$$\omega_l = \Phi(\alpha_l) \prod_{r < l} (1 - \Phi(\alpha_r)); \alpha_l \sim N(\mu_\alpha, \tau_\alpha) \quad (13)$$

we introduced the latent variable s_i and $z_{il}(s_i)$, $s_i = l$ denoted the i -th position was assigned to the l -th component,

$$s_i \sim \text{multinomial}(p_1, \dots, p_{L-1}) \quad (14)$$

$$z_{il}(s_i) = \begin{cases} N^-(\alpha_l(s_i), 1) & l < s_i \\ N^+(\alpha_l(s_i), 1) & l = s_i \end{cases} \quad (15)$$

For component specific parameters μ_l, σ_l , we assumed conjugate normal and gamma hyperpriors:

$$\mu_l \sim N(\mu_\mu, \tau_\mu); \tau_l \sim \text{Gamma}(a_\tau, b_\tau) \quad (16)$$

For variable selection parameter γ_i , we assumed a Bernoulli-Beta conjugate prior:

$$\gamma_i \sim \text{Ber}(\kappa); \kappa \sim \text{beta}(a_0, b_0) \quad (17)$$

Assuming data were properly normalized and standardized, we adopted the following choices for the hyperparameters. For $G(\cdot)$, $\mu_l = \{-5, -2, 0.1, 1, 2\}$; $\tau_l = \{0.4, 1, 1, 1, 1\}$; $\mu_\alpha = 0$; $\tau_\alpha = 1$; $\mu_\mu = 0$; $\tau_{mu} = 1$; $a_\tau = b_\tau = 0.5$. For variable selection prior, $a_0 = b_0 = 0.5$.

C.3 MCMC ALGORITHM

Step 1. Update s_i for $i = 1, \dots, m$, given $\gamma_i = 0$

$$Pr(S_i = l) = \frac{\omega_l N(y_i | \mu_l, \tau_l)}{\prod_{l=1}^L \omega_l N(y_i | \mu_l, \tau_l)} \quad (18)$$

where $\omega_l = \Phi(\alpha_l) \prod_{r < l} (1 - \Phi(\alpha_r))$.

Step 2. Update z_{il} for $i = 1, \dots, m$ and $l = 1, \dots, L$

$$z_{il}(s_i) = \begin{cases} N^-(\alpha_l(s_i), 1) & l < s_i \\ N^+(\alpha_l(s_i), 1) & l = s_i \end{cases} \quad (19)$$

Step 3. Update α_l for $l = 1, \dots, L$,

$$\alpha_l \sim N\left(\frac{\sum_{S_i > l} z_{il} + \mu_\alpha}{n_l + 1}, \frac{1}{n_l + 1}\right) \quad (20)$$

where $n_l = \sum_{i=1}^m 1(S_i \geq l)$.

Step 4. Update μ_l for $l = 1, \dots, L$,

$$\mu_l \sim N\left(\frac{\mu_\mu \tau_\mu + \tau_l \sum y(S_i = l)}{\tau_\mu + n_2 \tau_l}, \tau_\mu + n_2 \tau_l\right) \quad (21)$$

where n_2 is the number of elements in $y(S_i = l)$.

Step 5. Update τ_l for $l = 1, \dots, L$,

$$\tau_l \sim \text{Gamma}\left(a_\tau + n_2, b_\tau + 0.5 \sum (y - \mu_l)\right) \quad (22)$$

Step 6. Update γ_i for $i = 1, \dots, m$

$$Pr(\gamma_i = 1) = \frac{a_i}{a_i + b_i} \quad (23)$$

$$a_i = \kappa \int \prod N(y(S_i = l) | \mu_l, \tau_l) f(\mu_l) f(\tau_l) d\mu_l d\tau_l \quad (24)$$

$$b_i = (1 - \kappa) \prod N(y | \mu_0, \tau_0) \quad (25)$$

where $f(\mu_l)$ and $f(\tau_l)$ are distributions of μ_l and τ_l , respectively, the integration in equation (17) could be solved using the approximation:

$$p(y) = \frac{p(y|\mu_l, \tau_l)f(\mu_l)f(\tau_l)}{p(\mu_l|y)p(\tau_l|y)} \quad (26)$$

Step 7. Update κ

$$\kappa \sim \text{Beta}(a_0 + \sum \gamma_i, b_0 + m - \sum \gamma_i) \quad (27)$$

C.4 SNP ARRAY AND WES DATA PROCESSING

In the application of BMI-CNV to analyze 1000 genomes project and HapMap datasets (D. M. Altshuler *et al.*, 2010; Auton, Gonalo R. Abecasis, *et al.*, 2015). all Affymetrix raw CEL files were downloaded from website (<ftp://ftp.ncbi.nlm.nih.gov/hapmap>). We used the Affymetrix Power Tools and PennCNV package to generate LRR signals (Wang *et al.*, 2007). For WES data, the raw BAM files were downloaded from website (<ftp://ftp.1000genomes.ebi.ac.uk>). The BAM files were processed, sorted and filtered using SAMtools to generate raw read count (Li *et al.*, 2009). We then used the three-step normalization procedure described in Section A.1 to calculate the $\log_2 R\text{-}MED$ data, 4 external samples: NA10851, NA18502, NA12272, NA19072 were selected as control samples.

In the application of BMI-CNV to analyze samples from international lung cancer study (TRICL) with both OncoArray data and WES data. (Amos *et al.*, 2017). The OncoArray was designed from a list of 533,000 SNP markers. To retain high-quality genotype data, we applied the following quality control (QC) filters to remove (1) low quality samples (call rate<0.95); (2) unexpected duplicated and related samples (identical by descent (IBD)>0.2). Intensity data was obtained for each probe using GenomeStudio, genomic wave was adjusted by PennCNV. For WES data, QC procedures including base call quality recalibration variant filtering, genotypes refinement and Principal component analysis (PCA) of quality metric to exclude quality outliers. Kinship coefficient was also calculated to identify and exclude duplicated and related samples (Zheng *et al.*, 2012). Raw read count data were then generated and normalized.

C.5 POST-CALLING CNV QUALITY CONTROL (QC)

After obtaining the raw CNVs, to retain high quality CNV call, we first merged adjacent CNV calls (<5 markers). We then applied the following CNV pruning and filtering procedures including removing CNV calls that were (1) overlapped with centromeric regions; (2) < 3 exons or <10 SNPs, >100 markers and >500kb.

C.6 NUMERICAL SIMULATION

DATA SIMULATION

For SNP array data, we simulated the log R ratio (LRR) and B allele frequency (BAF) from the normal distribution,

$$LRR \sim N(\mu_{LRR}, \sigma_{LRR}^2) \quad (28)$$

$$BAF \sim N(\mu_{BAF}, \sigma_{BAF}^2) \quad (29)$$

The empirical mean and variance values were provided by Illumina website (<https://www.illumina.com/documents>) and summarized below (Table A.6.1).

For WES data, we spiked in the CNV signals by multiplying the raw read depth by a factor $c/2$, c was sampled from normal distribution,

$$c \sim N(\mu_{CNV}, \sigma_{CNV}^2) \quad (30)$$

Where the choices of mean and standard deviation (i.e. μ_{CNV}, σ_{CNV}) parameters of different copy number states were summarized below (Table A.6.2).

PERFORMANCE EVALUATION METRICS

The performance of all the calling methods was assessed by precision rate, recall rate and F1 score measures. The precision rate which measured the proportion of CNV calls from methods that overlapped with the true CNV set was defined as, True positives/(True positives + False positives), while the recall rate which measured the proportion of true CNVs that were called by methods was defined as, True positives/(True positives + False

negatives). The F1 score was defined as the harmonic mean of precision and recall rate

which reflected the overall accuracy, $2 \frac{precision * recall}{precision + recall}$.

Table C.1 Copy number states and empirical parameter settings for SNP array data.

Copy Number State	LRR mean	LRR SD	BAF mean	BAF SD
Double Deletion	-5.00	2.00	NA	NA
Single Deletion	-0.45	0.18	0, 1	0.03
Normal	0.00	0.18	0, 0.5, 1	0.03
Single Duplication	0.30	0.18	0, 1/3, 2/3, 1	0.03
Double Duplication	0.75	0.18	0, 1/4, 2/4, 3/4, 1	0.03

LRR: log R ratio; BAF: B allele frequency; SD: standard deviation.

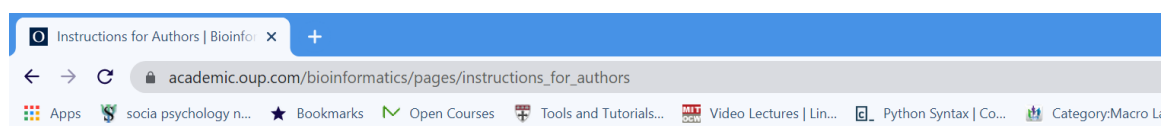
Table C.2 Copy number states and empirical parameter settings for WES data.

Copy Number State	μ_{CNV}	σ_{CNV}
Double Deletion	0.05	0.10
Single Deletion	1.00	0.10
Single Duplication	3.50	0.10
Double Duplication	6.00	0.10

APPENDIX D

PERMISSION TO REPRINT

D.1 COPYRIGHT RELEASE FROM JOURNAL *BIOINFORMATICS*.

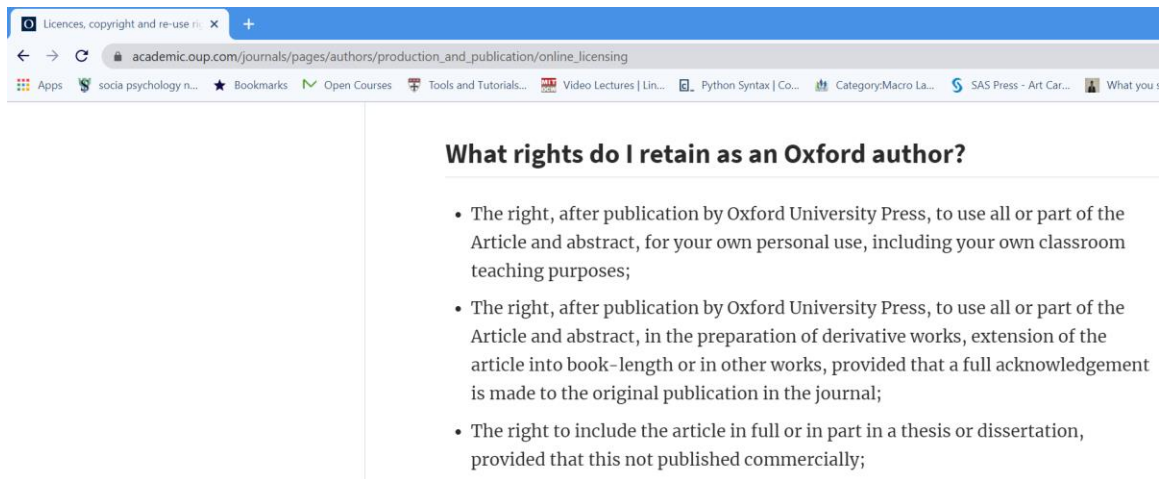


Because accepted manuscripts are published online following acceptance, it is important that the final version of the manuscript supplied by the author contains no information regarding the citation information (volume, issue, year) or a copyright line as this will mislead readers.

Licence to Publish

It is a condition of publication in Bioinformatics that authors grant an exclusive licence to Oxford University Press. This ensures that requests from third parties to reproduce articles are handled efficiently and consistently and will also allow the article to be as widely disseminated as possible. As part of the licence agreement, Authors may use their own material in other publications provided that Bioinformatics is acknowledged as the original place of publication and Oxford University Press as the Publisher. [Information about the Creative Commons licence.](#)

D.2 COPYRIGHT RELEASE FROM JOURNAL *BRIEFINGS IN BIOINFORMATICS*.



The screenshot shows a web browser window with the address bar displaying `academic.oup.com/journals/pages/authors/production_and_publication/online_licensing`. The page title is "Licences, copyright and re-use". The main heading is "What rights do I retain as an Oxford author?". Below the heading, there is a list of three bullet points detailing the rights retained by an author after publication by Oxford University Press.

What rights do I retain as an Oxford author?

- The right, after publication by Oxford University Press, to use all or part of the Article and abstract, for your own personal use, including your own classroom teaching purposes;
- The right, after publication by Oxford University Press, to use all or part of the Article and abstract, in the preparation of derivative works, extension of the article into book-length or in other works, provided that a full acknowledgement is made to the original publication in the journal;
- The right to include the article in full or in part in a thesis or dissertation, provided that this not published commercially;