DETECT COPY NUMBER VARIATIONS FROM READ-DEPTH OF HIGH-THROUGHPUT SEQUENCING DATA

Weibo Wang

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Computer Science.

Chapel Hill
2015

Approved by:

Wei Wang

Wei Sun

Jin Szatkiewicz

Leonard McMillan

Jan Prins

ABSTRACT

Weibo Wang: Detect Copy Number Variations from Read-depth of High-throughput Sequencing Data
(Under the direction of Wei Wang)

Copy-number variation (CNV) is a major form of genetic variation and a risk factor for various human diseases, so it is crucial to accurately detect and characterize CNVs. High-throughput sequencing (HTS) technologies promise to revolutionize CNV detection but present substantial analytic challenges. This dissertation investigates improving the CNV detection using HTS data mainly from the following aspects.

- It is observed that various sources of experimental biases in HTS confound read-depth estimation, and bias correction has not been adequately addressed by existing methods. This dissertation presents a novel read-depth-based method, GENSENG, which identify regions of discrete copy-number changes while simultaneously accounting for the effects of multiple confounders.
- It is conceivable that allele-specific reads from HTS data could be leveraged to both enhance CNV detection as well as produce allele-specific copy number (ASCN) calls. Although statistical methods have been developed to detect CNVs using whole-genome sequence (WGS) and/or whole-exome sequence (WES) data, information from allele-specific read counts has not yet been adequately exploited. This dissertation presents an integrated method, called AS-GENSENG, which incorporates allele-specific read counts in CNV detection and estimates ASCN using either WGS or WES data.
- Although statistically powerful, the GLM+NB method used in GENSENG and AS-GENSENG has a quadric computational complexity and therefore suffers from slow running time when applied to large-scale sequencing data. This dissertation aims to

substantially speed up the GLM+NB method by using a randomized algorithm and demonstrate the utility of our approach by providing R-GENSENG, a speeded up version of GENSENG.

To my family and advisor, I couldn't have done this without you.
Thank you for all of your support along the way.

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratefulness to my advisor Dr. Wei Wang for her continuous guidance and support. I feel especially fortunate to have worked closely with Dr. Jin Szatkiewicz and Dr. Wei Sun who encouraged me to work persistently, and greatly helped me to improve my critical thinking ability and writing skill. Special thanks go to Dr. Jan Prins who chaired my committee and provided assistance to me. I would also like to thank Dr. Leonard McMillan who served on my committee and devoted a lot of effort to my study.

My special thanks also go to members in CompGen Lab, including Eric Yi Liu, Zhaojun Zhang, Shunping Huang, Wei Cheng, etc., for their thoughtful discussions on the problems in the research. I would like to thank all fellow persons I met in the Computer Science department who provide all kinds of help to me during my entire pursuit of PhD. I would also like to thank all the warm-hearted persons I met in the past few years for helping me to live here in Chapel Hill, the beautiful town in North Carolina.

Finally, I am also deeply thankful to my parents. They stand steadily after me, encourage me to overcome the difficulties I faced in my pursuit of PhD. Without their endless support, I have absolutely no chance to finish this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Background

1.1.1 Significance of Detecting Copy Number Variation

Copy-number variants (CNVs) are a major form of genetic variation in mammals (Clop et al., 2012; Conrad et al., 2010; Mills et al., 2011; Yalcin et al., 2011). CNVs have been shown to affect gene expressions in human cell-lines (Stranger et al., 2007) as well as in different tissues of rodents (Cahan et al., 2009; Guryev et al., 2008; Henrichsen et al., 2009), and to play an important role in the etiology of schizophrenia (Sklar et al., 2008; Stefansson et al., 2008), autism (Malhotra and Sebat, 2012; Sebat et al., 2007), and non-psychiatric diseases (Bochukova et al., 2010; Fanciulli et al., 2007; Walters et al., 2010). Indeed, CNV assessment is beginning to become a routine part of the diagnostic workup for some medical conditions, including neurobehavioral disorders (Levinson et al., 2011; Doco et al., 1990; Cooper et al., 2011; Sullivan et al., 2012). CNV assessment is also important in functional genomic studies since failing to account for copy-number differences can result in misinterpretation of data from RNA-seq, chromatin immunoprecipitation (ChIP-seq), DNase-hypersensitive site mapping (DNase-seq), or formaldehyde-assisted isolation of regulatory elements (FAIRE-seq) (Laird, 2010; Rashid et al., 2011a). For these reasons, accurate detection of CNVs is of paramount importance; and allele-specific copy number (ASCN) calls are highly desirable as it is also important to know how CNVs are allocated in diploid organisms(Tewhey et al., 2011; Gamazon et al., 2014). For example, allele-specific copy number analysis of breast tumors allowed the construction of a genome-wide map of allelic skewness in breast cancer (Van Loo et al., 2010). Furthermore, many recessive Mendelian disorders, such as Cohen syndrome (Balikova et al., 2009), often result from the unmasking of a deleterious allele by a one copy deletion. Therefore, allele-specific CNV calls provide crucial additional information for disease

studies. Using genome-wide SNP arrays (Attiyeh et al., 2009; Gardina et al., 2008; Greenman et al., 2010; Pounds et al., 2009), allele-specific intensity signals for two SNP alleles (denoted as alleles A and B) can be obtained and integrated in CNV detection. ASCN calls can then be generated (e.g., A, AAB, BBB, ABBB). ASCN calls provide a more accurate characterization of the underlying DNA sequence of each individual, thereby reducing the rate of apparent Mendelian inconsistencies (McCarroll et al., 2006; Korn et al., 2008) and could improve statistical power for tests of association with complex diseases(Marenne et al., 2013).

1.1.2 CNV Detection Methods based on High Throughput Sequencing Data

Microarray technologies were the main platform for initial work in CNV characterization (Iafrate et al., 2004; Sebat et al., 2004) and remain a cost-efficient choice (Alkan et al., 2011). The array comparative genomic hybridization (CGH) technology requires paired samples (test and genome), and calls CNV based on checking the signal ratio between the test and the genome at the pre-defined set of targets. The SNP microarray technology performs hybridization on one sample and in addition to the intensities signal measured at probes across samples, it also provides a feature called B allele frequency (BAF), which could increase the CNV detection sensitivity. With the latest advancing in both array CGH and SNP microarray technologies, the smallest CNV could be detected has been refined to 500 bps (Alkan et al., 2011). To call ASCN, several methods have been developed for CNV detection using allele-specific probe intensities from SNP arrays (Korn et al., 2008; Wang et al., 2007; Sun et al., 2009). With Affymetrix array data, Birdsuite uses a hidden Markov model (HMM) and defines allele-specific properties of each probe through HMM emission probability(McCarroll et al., 2006). With Illumina array data, raw intensity data is transformed to the total intensity from both alleles (i.e. log R Ratio or LRR) and the relative ratio of the intensity between two alleles (i.e., B Allele Frequency or BAF). HMM-based methods, such as PennCNV (Wang et al., 2007) and GenoCN (Sun et al., 2009), jointly analyse LRR and BAF in the likelihood. According to simulations and studies on individuals with known CNVs, integrating allele-specific information in array-based CNV calling not only yields ASCN, but also it improves the accuracy of total copy-number calls (e.g., 1 copy deletion, 3 copy duplication).

Recent advances in high-throughput sequencing (HTS) (Bentley et al., 2008; McKernan et al., 2009; Wheeler et al., 2008) are promoting whole-genome sequencing (WGS) or whole-exome sequencing (WES) as an all-in-one high-throughput assay for characterizing single-nucleotide polymorphisms (SNPs) and CNVs, and could replace microarrays as a discovery platform (Alkan et al., 2011; Medvedev et al., 2009). While microarray-based CNV detection analyzes probe hybridization intensities, HTS-based CNV detection uses conceptually distinctive approaches: read-pair, split-read, and read-depth analyses, which vary in their sensitivity and specificity depending on the sizes and classes of SVs (Mills et al., 2011; Alkan et al., 2011; Medvedev et al., 2009). Converging evidence suggests that multiple approaches should be considered together to maximize CNV detection from HTS data. For example, the 1000 Genomes Project used 19 algorithms to independently identify CNVs in 185 human genomes and pooled the results according to the specificity of each algorithm (Mills et al., 2011). Recent methods (SPANNER (Mills et al., 2011), CNVer (Medvedev et al., 2009) and Genome STRiP (Handsaker et al., 2011)) integrate read-pair and read-depth in the detection process in different ways.

The read-depth approach looks for higher or lower than expected sequencing coverage in a genomic region to infer gain or loss of DNA. Read-depth has been computed in a variety of ways, including counting the number of fragments (Handsaker et al., 2011) or reads (Abyzov et al., 2011; Campbell et al., 2008; Yoon et al., 2009; Medvedev et al., 2010) mapped to a particular genomic region and calculating the sum of per-base coverage within a region (Simpson et al., 2009; Sudmant et al., 2010). Existing CNV detection methods assume that read-depth follows a Poisson distribution (or a normal distribution as the large-sample approximation of the Poisson model) for a diploid genome and search for regions that diverge from this distribution. However, in practice, neither sampling nor mapping of the reads is uniform, because of experimental biases. GC content can lead to certain genomic regions being over- or under-sampled (Bentley et al., 2008). Repetitive DNA elements are abundant in the mammalian genomes (Treangen and Salzberg, 2012); consequently, the number of reads unambiguously mapped to a region could be very different from the number of reads sequenced from the region. Additional sources of bias, which are more difficult

to trace (e.g., noise arising from sequencing, sequencing errors), create further variability in read-depth coverage. Violation of the assumed Poisson distribution entails loss of sensitivity/specificity to detect CNVs using read-depth. Thus, when analysing HTS data, it is critical to correct for various sources of experimental bias that distort the quantitative relationship between read-depth and true copy number, hindering the ability for accurate CNV detection (Bentley et al., 2008; Treangen and Salzberg, 2012; Szatkiewicz et al., 2013).

Bias correction has not been adequately addressed in the literature. In studies of cancer, matched pairs of tumor- and normal-tissue samples may be used to correct biases by computing read-depth ratios (Chiang et al., 2009; Xi et al., 2011; Xie and Tammi, 2009; Ivakhno et al., 2010). While cancer studies afford themselves the use of tumour/normal pairs, numerous techniques have been developed for germline CNV studies to normalize read-depth data (a.k.a. correct bias) (Handsaker et al., 2011; Abyzov et al., 2011; Yoon et al., 2009; Medvedev et al., 2010; Simpson et al., 2009; Sudmant et al., 2010; Chiang et al., 2009; Xi et al., 2011; Xie and Tammi, 2009; Wang et al., 2013). For WGS, most existing methods (Abyzov et al., 2011; Yoon et al., 2009; Sudmant et al., 2010) use a two-step approach, where read-depth data from a single-genome is first adjusted to account for the effect of known sources of bias (e.g. GC content) and then the adjusted read-depth is segmented to predict CNVs. While various kinds of adjusted read-depth have been used as input, nearly all methods (Handsaker et al., 2011; Yoon et al., 2009; Medvedev et al., 2010; Simpson et al., 2009) employ the Poisson or normal-distribution assumption without subsequent evaluation the adequacy of the distributional assumption. Exome sequencing introduces additional sources of noise to the raw read-depth data (Fromer et al., 2012; Amarasinghe et al., 2013; Coin et al., 2012; Karakoc et al., 2012; Koboldt et al., 2012; Nord et al., 2011; Plagnol et al., 2012; Krumm et al., 2012; Li et al., 2012; Love et al., 2011; Wu et al., 2012; Tan et al., 2014) and methods developed for WES data typically leverage the large-scale nature of exome sequencing projects for noise-reduction/data-normalization. Based on their noise-reduction techniques, most existing WES methods can be classified into two categories: either multivariate methods including principle component analysis (PCA) and singular value decomposition (SVD), or reference-set

methods (Amarasinghe et al., 2013; Nord et al., 2011; Plagnol et al., 2012; Li et al., 2012; Love et al., 2011; Wu et al., 2012; Tan et al., 2014). The PCA/SVD methods assume that most variation observed in the sample-by-target read-depth matrix is due to noise with little contribution from CNVs; and therefore remove several of the strongest variance components for the purpose of noise reduction. In this paradigm, XHMM (Fromer et al., 2012) applies a PCA that is optimized for detecting rare CNVs (frequency $< 5\%$), whereas common CNVs could not fit in this model. CoNIFER (Krumm et al., 2012) applies a SVD and removes the first 12-15 variance components for detecting rare CNVs but 5 components for common CNVs. However, as the frequencies of CNVs cannot be known before they are detected, it is challenging to determine how to choose the top-K variance components in order to prevent the PCA/SVD methods from removing true CNV signals (Tan et al., 2014). Alternatively, the reference-set methods create a baseline for each exon target from a reference group of copy number two, where the baseline from the reference set captures technical variation but not variation due to CNVs. Then read-depth ratios of test samples versus the baseline are computed for the purpose of noise reduction (Amarasinghe et al., 2013; Nord et al., 2011; Plagnol et al., 2012; Li et al., 2012; Love et al., 2011; Wu et al., 2012). However, the power to detect common CNVs is often limited, owing to the difficulty in constructing the true reference set in the presence of common CNVs, especially when the CNV frequency is high and unknown (Plagnol et al., 2012; Li et al., 2012). Here we demonstrate that allele-specific read count information can be leveraged to identify the proper reference group of samples with copy number two and this method subsequently improves detection of common CNVs at any frequency.

Analogous to microarray-based methods, allele-specific information could also be leveraged for HTS-based CNV detection. For cancer studies, specific methods (Patchwork (Mayrhofer et al., 2013), SomatiCA (Chen et al., 2013), WaveCNV (Holt et al., 2014), ADTEx (Amarasinghe et al., 2014)) have been developed to incorporate allele-specific information in detecting copy number aberration using tumour/normal sample pairs. Such methods typically apply a two-step approach, where read-depth ratios of the tumour/normal pairs and the minor allele frequency data are analysed separately. However, for detecting germline CNVs, allele-specific

information has not been extensively explored in the literature. With WGS data, ERDS (Heinzen et al., 2012) is the only existing method that leverage allele-specific information but it has a number of limitations. For example, deletions are detected by simultaneous analysis of read-depth and the total number of heterozygous SNPs followed by refinement of smaller segments ($<$10kb) using read-pair information; however, duplications are detected using read-depth only. Further, ERDS estimates total copy-numbers but it is not capable of estimating ASCN. For WES data, many effective methods have been developed to estimate rare or common CNVs (Fromer et al., 2012; Amarasinghe et al., 2013; Coin et al., 2012; Karakoc et al., 2012; Koboldt et al., 2012; Nord et al., 2011; Plagnol et al., 2012; Krumm et al., 2012; Li et al., 2012; Love et al., 2011; Wu et al., 2012); however, none of the existing methods leverage allele-specific information in CNV detection or are capable of estimating ASCN. To overcome these deficiencies, here we develop a novel method that uses allele-specific information to aid the detection of both deletions and duplications and is capable of determining ASCN from both WGS and WES data.

1.1.3 Computational Burden in the Read-count based CNV Analysis

The read-depth data (windowed or gene/exon-level read counts) are by nature a series of counts, for which the negative-binomial (NB) distribution has been shown as the suitable distribution in statistical modeling (Robinson and Smyth, 2008; Anders and Huber, 2010; Rashid et al., 2011b; Szatkiewicz et al., 2013). The NB model is flexible for modeling genomic read-count data because its dispersion parameter allows larger variance and therefore less restrictive than Poisson distribution. Further, via generalized linear models (GLMs) (McCullagh, 1983), the NB model provides a powerful framework to simultaneously account for confounding factors (e.g. genomic GC content and mappability) and determine the true relationships between read-count signals and biological factors (Szatkiewicz et al., 2013).

A large number of statistical methods and software tools have been developed to couple GLM and NB models (referred to as GLM+NB hereafter) for analyzing genomic read-count data. For example, GENSENG (Szatkiewicz et al., 2013) was developed for detecting CNVs using DNA-seq; DESeq (Anders and Huber, 2010) and edgeR (Robinson et al., 2009; McCarthy et al.,

2012; Robinson and Smyth, 2007, 2008; Zhou et al., 2014) for detecting differential expression using RNA-seq; and ZINBA (Rashid et al., 2011b) for detecting enriched regions using ChIP-seq, DNase-seq, or FAIRE-seq. While statistically powerful, GLM+NB methods encounter a big data problem when applied to genomic read-count data comprised of tens of thousands of windows/genes/exons. The iterative reweighed least square (IRLS) algorithm is the standard approach to fit GLMs (Green, 1984). IRLS has a quadric complexity with the size of data and needs to be run multiple times until it converges. The expensive computation cost of GLM hinders the computational efficiency of the GLM+NB methods in their applications to genomic read-count data.

1.2 Thesis Statement

Thesis:*Accurate and efficient CNV detection based on the analysis of read-depth of HTS data could be achieved by correcting biases in HTS and leveraging the allele specific information accompanied with HTS data.*

This dissertation presents an integrated likelihood-based CNV inference framework, which is based on Hidden Markov Model (HMM), that combines multiple information along with the read-depth in HTS data to detect CNV. Three specific design aims of the framework are listed below:

- ***Specific Aim 1:****Improve CNV detection accuracy through correcting biases.* Biases could lead to read-depth diverges from expected values and cause false positive CNV calls. However, the effect of biases could be quantified and corrected so that read-depth could be correctly analyzed to achieve more accurate CNV detection.
- ***Specific Aim 2:****Allele specific information could not only be used to generate ASCN, but also it could lead to accurate CNV predictions.* ASCN is highly desirable. The success of CNV detection in microarray study inspires the incorporation of the allele specific information in HTS based CNV detection.
- ***Specific Aim 3:****The CNV prediction could be both accurate and efficient.* The computational cost could be highly expensive due to the large volume of HTS data and the computational complexity in the sophisticate GLM+NB methods. Nevertheless, the

computation efficiency could be improved by randomly sampling a subset of data before the analysis so that the scale of analysis is reduced. Even so, the accuracy should still be comparable with using all data.

As the support of the thesis statement and the research aims above, the contributions made in the thesis are summarized as follows.

- A method GENSENG is developed that simultaneously corrects biases and segment CNV. HMM is employed to respect the spatial nature of the CNV (consecutive regions tend to have the same copy number). Through evaluation of read-depth data distribution using both human data and mouse data, negative-binomial distribution is found as a better distribution because it captures the characteristics of real data by allowing the variance larger than the mean. A negative-binomial regression is applied, which is a specific case of the generalized linear model (GLM), to quantify the relation between read count, underlying copy number, and biases (GC content and mappability). The negative-binomial regression procedures emission probability of HMM. HMM generates the posterior probability given each underlying copy number for each window, and based on the posterior probability GENSENG segments CNV. GENSENG achieves higher CNV detection accuracy than its peer methods which do not adequately quantify the effect of biases in terms of sensitivity and FDR.
- The allele-specific information is incorporated into the likelihood-based framework in GENSENG. The extended framework, which is called AS-GENSENG, models the allele-specific information captured in allele-specific reads using the binomial distribution. The joint likelihood of a given copy number from both the read-depth information and the allele-specific information is used to define the emission probability in AS-GENSENG. AS-GENSENG further predicts ASCN through collecting the allele specific information inside the HMM segmented CNV regions. For WES data where the frequency of CNV is unknown and might be high, AS-GENSENG utilizes the allele-specific information to estimate accurate expected read count without assumption on the CNV frequency.

AS-GENSENG brings ASCN and is capable of detecting CNV from both WGS and WES data. It further improves the CNV detection accuracy over GENSENG and has better accuracy than its WGS companion methods when the size of CNV is larger than 1kbps. With the accurate expected read count estimation AS-GENSENG achieves higher accuracy than other WES methods when the CNV frequency is high (common CNV).

- In this study, we introduce the randomized GLM+NB coefficients estimator (RGE) for speeding up the GLM+NB based read-count analysis. Our RGE uses a weighted sampling strategy. To illustrate the utility of RGE, we used our GLM+NB based CNV detection method GENSENG (Szatkiewicz et al., 2013) as an example and named the resulting RGE-GENSENG as "R-GENSENG". We first evaluated the consistency and the variance properties of RGE. We concluded that RGE is a consistent GLM+NB regression estimator, and that the weighting sampling strategy applied in RGE yields smaller regression coefficients estimation variance than using uniform sampling. We then performed simulation and real-data analysis to evaluate R-GENSENG in comparison to the original GENSENG. We concluded the R-GENSENG is ten times faster than the original GENSENG while maintaining GENSENGs accuracy in CNV detection. Taken together, our results suggest that RGE and the strategy developed in this work could be applied to other GLM+NB based read-count analyses in order to substantially improve their computational efficiency while preserving the analytic power.

1.3 Summary

In this chapter, I have illustrated the extreme importance of accurate CNV detection and ASCN profiling. With the advance of HTS technology, it is highly demanding to develop computational tools to accurately and efficiently analyze the huge amount of HTS data generated every day. In summary, the goal of this dissertation is to develop efficient methods that not only predicts CNV and ASCN for both WGS and WES data, but also predicts with high accuracy. The remaining chapters are organized as follows:

- In Chapter 2, I will present the detailed descriptions of GENSENG method. GENSENG has been validated extensively on both simulation data and real data. I will introduce the simulation procedures as well as the sources of real data in Chapter 2. I will also introduce the experiments setup and the comparison results of GENSENG with the state of the art methods.
- In Chapter 3, I will present the method characterization of AS-GENSENG. The evaluation of AS-GENSENG is conducted on two different data, WGS data and WES data, respectively, and for each data, both simulation study and real data study are conducted. I will introduce details of the conducted experiment results in Chapter 3. Through extensive comparison with peer methods, AS-GENSENG is concluded to not only predict ASCN, but also improve the CNV detection accuracy. I will introduce the comparisons and bring the discussions.
- In Chapter 4, I will introduce the RGE and the efficiently implemented R-GENSENG. I will first show the properties of the RGE, and show its application R-GENSENG. The properties of RGE are established by both theoretic derivation and simulation validation, which are both covered in Chapter 4. The performance of R-GENSENG on accurate and efficient CNV detection is evaluated on both simulation data and real data. I will present the results in Chapter 4.
- In Chapter 5, I will conclude the dissertation work and show future directions to achieve further improvement based on the work in this dissertation.

CHAPTER 2: GENSENG

2.1 Overview

In Section 1, we have mentioned that bias correction has not been adequately addressed in the literature. Experimental biases would seriously affect the read-depth and leads to false CNV calls. Some existing methods (Abyzov et al., 2011; Yoon et al., 2009; Sudmant et al., 2010) adopt a two-step approach where read-depth is first smoothed for GC content differences using linear regression, and the GC-adjusted read-depth is then segmented. Other methods (Handsaker et al., 2011; Medvedev et al., 2010) account for mapping bias of a candidate region by using its effective length (e.g. the number of confidently mapped bases); however, this approach does not account for the dependence between consecutive regions or additional sources of noise in the data. While various kinds of adjusted read-depth have been used as input, nearly all methods (Handsaker et al., 2011; Yoon et al., 2009; Medvedev et al., 2010; Simpson et al., 2009) employ the Poisson or normal-distribution assumption without subsequent evaluation the adequacy of the distributional assumption.

In this chapter, we first show that evaluation of the distribution assumption for read-depth is important, as it may not hold true. Second, we show that it is important to jointly estimate copy number and the effect of confounding factors. Third, we develop a novel statistical method to accurately model read-depth and detect CNVs from HTS data. We measure read-depth by the number of sequence fragments mapped in sliding windows tiled along the genome; and we model the fragment counts by negative binomial distributions, which allow for over-dispersion and account for the effects of confounders. Furthermore, we account for the dependence of fragment counts of adjacent windows by employing a hidden Markov model (HMM). Known confounding factors are treated as covariates and corrected explicitly, while unknown experimental biases are

accommodated by the over-dispersion parameter of the negative binomial distribution and by an additional noise component via a mixture model. Fourth, we calibrate our method using simulation and whole-genome-sequencing data from the 1000 Genomes Project (1000GP) (Mills et al., 2011); and we compare our method to CNVnator (Abyzov et al., 2011), the best-performing read-depth-based CNV detection algorithm in the literature (Mills et al., 2011). Finally, to demonstrate the utility and robustness of our method, we apply our method to both human and mouse HTS datasets.

In summary, our method outperforms existing read-depth-based CNV detection algorithms and distinguishes homozygous and heterozygous deletions and high-copy duplications. Our method complements the current literature, and the concept of simultaneous bias correction and CNV detection can serve as a basis for combining read-depth with read-pair or split-read in a single analysis. A user-friendly and computationally efficient implementation of our complete analytic protocol is freely available at `https://sourceforge.net/projects/genseng/`.

2.2 Materials and Methods

In this section, the studied sequencing datasets are first introduced then followed by the procedure to convert them to read-depth data. After that, a developed CNV detection method GENSENG (Szatkiewicz et al., 2013) is introduced.

2.2.1 Datasets included in this study

1000 Genome Project data. For method development and assessment, we used the whole-genome sequencing data from 3 HapMap individuals sequenced as part of the 1000 Genomes Project. These include the CEU parent-offspring trio of European ancestry (NA12878, NA12891, NA12892), sequenced to 42x coverage on average using the Illumina Genome Analyzer (I and II) platform. Sequencing reads were a mixture of single-end and paired-end with variable lengths (36bp, 51bp). The complete genome sequence data were obtained in the form of .bam alignment files from `ftp://ftp.ncbi.nlm.nih.gov/1000genomes/ftp/pilot_data/data/`. Reads were aligned to the human reference genome NCBI37 using BWA ((Li and Durbin, 2009)) (v0.5.5)

as described in the on-line documentation:
`ftp://ftp.ncbi.nlm.nih.gov/1000genomes/ftp/README.alignment_data`.

High-confidence CNVs To assess the sensitivity of our discovery method, we used the high-confidence CNVs established for NA12878 (Supplementary Table 6 of (Mills et al., 2011)) by combining CNVs reported in earlier surveys that used high-density microarrays (Conrad et al., 2010; McCarroll et al., 2008), fosmid sequencing (Kidd et al., 2008), or ABI tracing mapping (Mills et al., 2006). This dataset included 610 deletions (∼82% from microarray reports (Conrad et al., 2010; McCarroll et al., 2008)) and 261 duplications (100% from microarray reports (Conrad et al., 2010; McCarroll et al., 2008)) from the autosomes of NA12878. The second high-confidence dataset used in this study was generated by Handsaker et al. (Handsaker et al., 2011) by accurately genotyping deletions from the 1000GP HTS data (Mills et al., 2011). The complete dataset was downloaded from
`ftp://ftp.broadinstitute.org/pub/svtoolkit/misc/1kg/NGPaper/` and included deletions for the 3 aforementioned HapMap individuals. This dataset included 2301 deletions for NA12878, 2200 deletions for NA12891, and 2055 deletions for NA12892. CNV coordinates reported in both high-confidence datasets were translated from NCBI36 to NCBI37 using liftOver.

Data from other sequencing projects. To demonstrate the robustness of our method, we used HTS data from two different studies using various sequencing platforms, sequencing depths, and read lengths. In the first study, we used the whole-genome sequencing data of 3 individuals affected with bipolar disorder from a large multiplex Spanish pedigree currently under investigation. Paired-end sequencing with 100bp reads was performed at the University of North Carolina on the Illumina HiSeq 2000 platform. Each individual was sequenced to an average of 15x coverage. Reads were aligned to the human reference genome NCBI37 using BWA (Li and Durbin, 2009) (v.0.5.5) with default parameters. In the second study, we downloaded (`ftp://ftp-mouse.sanger.ac.uk/current_bams/`) the whole-genome sequencing data of inbred mouse strains made freely available by the Mouse Genomes Project conducted at the

Sanger Institute (Keane et al., 2011). All mouse samples were sequenced on the Illumina GAII platform with a mixture of 54bp, 76bp, and 108bp paired reads to a coverage ranging from 17x to 43x. Reads were aligned to the mouse reference genome NCBI37 using the MAQ aligner (Yalcin et al., 2011; Li et al., 2008). For this study, we analyzed the alignment files for 13 inbred strains (129S1SvlmJ, A/J, AKR/J, BALB/cJ, C3H/HeJ, CAST/EiJ, CBA/J, DBA/2J, LP/J, NOD/LtJ, NZO/HILtJ, PWK/PhJ, WSB/EiJ). We also downloaded the released structural variation (SV) calls for these strains from `ftp://ftp-mouse.sanger.ac.uk/current_svs/`. These SVs have been classified into several categories based on specific paired-end mapping patterns briefly described in (Yalcin et al., 2011). From this SV release, we extracted 2 categories, including deletions and copy number gains (GAINS and TANDEMDUP), to compare to the GENSENG-predicted calls.

Reference genomes. The human reference genome NCBI37 was obtained from `ftp://ftp.ncbi.nlm.nih.gov/1000genomes/ftp/technical/reference/human_g1k_v37.fasta.gz`. The mouse reference genome NCBI37 was obtained from `ftp://ftp-mouse.sanger.ac.uk/ref/NCBIM37_um.fa`.

2.2.2 Input data preparation for CNV detection

Our input data is a triplet of read-depth (RD) signal, GC content, and mappability score computed in sliding windows tiled along the genome. RD signal: Alignment (.bam) files are parsed out using SAMtools (Li et al., 2009) with a quality control (QC) filter to extract confidently aligned reads. The procedure of the QC filters is as follows. 1. Remove any read that fails platform/vendor quality checks, or either a PCR duplicate or an optical duplicate. 2. Extract all single-end reads and properly paired paired-end reads. 3. Extract confidently aligned reads with MAPQ $\geq$ a specified threshold. In this study, we use MAPQ $\geq$10, which was empirically determined. A (single-end or paired-end) sequence read represents one or two ends of a DNA fragment randomly sampled from the donor/sample genome. Using reads passing the QC filter, the RD signal is calculated as the number of sequenced DNA fragments in sliding windows, ensuring each fragment is counted only once. Each read (e.g. 36-mer or 51-mer from the 1000 Genomes Project data (Abecasis et al., 2012;

Mills et al., 2011)) is represented by its middle base pair. A fragment is counted where read mapping information is available.

1. If two ends of a pair fall in two windows, assign 1/2 to each window where the ends fall;
2. If both ends of a pair fall in the same window, assign 1 to the window;
3. If paired-ended but only one-end present, assign 1/2 to the window where the ends fall;
4. If single-end, always assign 1 to the window where the end falls.

GC content: First we calculate the proportion of G or C bases in each window from a given reference genome. Then we apply a cubic spline smoothing and transform the GC proportion based on the fitted curve so that the transformed GC proportion and the logarithm of the read-depth are linearly correlated. Finally the transformed GC proportion is median-centered and is referred to as GC content hereafter.

Mappability score: As a function of both reference sequence and read length (K-mer), mappability score is calculated a priori in four steps: (1) Identify K-mers where each K-mer consists K consecutive bases starting at each base position from the reference genome. (2) Align the K-mers back to the reference genome using a desired aligner, e.g. BWA (Li and Durbin, 2009). Ideally, the aligner and the alignment parameters are chosen to match what was used for generating read alignment files from the sample genomes. (3) Identify mappable base positions where the corresponding K-mers map back to themselves unambiguously (i.e. there is a single best hit and it is the true position of the K-mer). For example, the X0 field produced by BWA (Li and Durbin, 2009) relates a K-mer from a specific place in the genome to the number of best hits of that K-mer in the entire genome. If a K-mer has a X0 value of 1, the corresponding base can be identified as a mappable base. (4) Compute mappability score as the proportion of mappable bases in a given window, which measures the uniqueness of specific regions of the reference genome.

Window consideration: The window size and the degree of overlap between them are adjusted for specific data. In this study, a window size of 500bp with 200bp overlap was chosen for all datasets for several reasons: First, the window size should be no less than the mean DNA fragment size of the sequencing library. Second, using a larger window size (e.g. 1Kb) or

non-overlapping windows would decrease precision in defining CNV breakpoints and miss CNVs that only partially span one window. Third, a higher degree of overlap introduces more inter-window correlation, which necessitates appropriate adjustment in modeling the RD signals.

In summary, the input data is a tuple for each window represent by

$\{O, X\} = \{o_1.....o_T, x_1.....x_T\}$, where T is the total number of windows of a chromosome, o_t denotes the read-depth, $x_t = (g_t, l_t)$ denotes the covariates of the t^{th} genomic region, where g_t represents the GC content, and l_t denotes the mappability score.

2.2.3 Overview of the GENSENG method

To mitigate the effects of experimental bias and improve read-depth-based CNV detection, we developed a novel statistical method called GENSENG (Szatkiewicz et al., 2013). The unique feature of GENSENG is to integrate the correction of multiple sources of bias and the inference of the copy-number states in a single analysis. Figure 2.1 gives an algorithmic overview of GENSENG. The required input contains two parts: the triplet data (read-depth, GC content and mappability score) and the initial parameter values. The input is passed to the GENSENG engine for parameter training based on the Baum-Welsh algorithm. To update the emission probability and the parameters for the negative binomial regression model, the weighted GLM fitting algorithm is applied iteratively, which uses the updated posterior probability of the copy-number state as the regression weights in each iteration. At the convergence of parameter training, GENSENG identifies the state with the largest posterior probability and assigns the associated copy number to the corresponding window. Finally, GENSENG outputs the coordinates of CNV segments and the confidence scores. The tasks of input preparation were implemented in R, perl, and python programming languages. The computational core of GENSENG was implemented in C++. Recommendation for the tuning parameters and initial emission parameters are provided as part of the software release. Given the input, GENSENG can report CNVs from a $\sim$40x human chromosome within a couple of hours.

The methodological details of GENSENG are described in Section 2.2.4. Briefly, the key components of GENSENG are summarized below. First, we used an HMM with seven states (0-6)

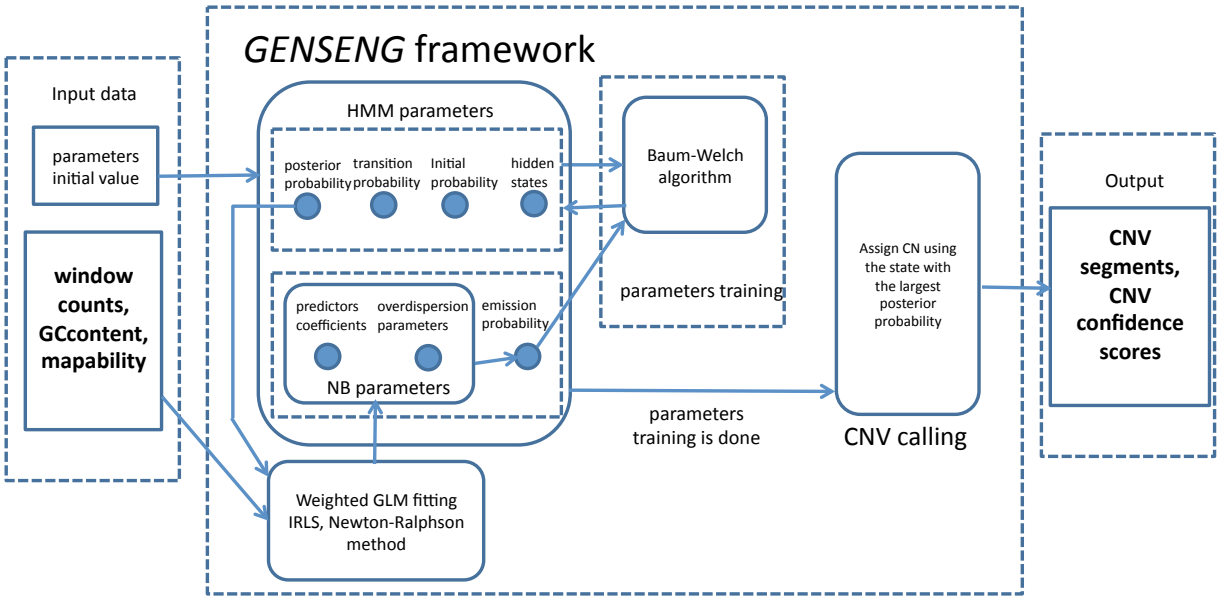
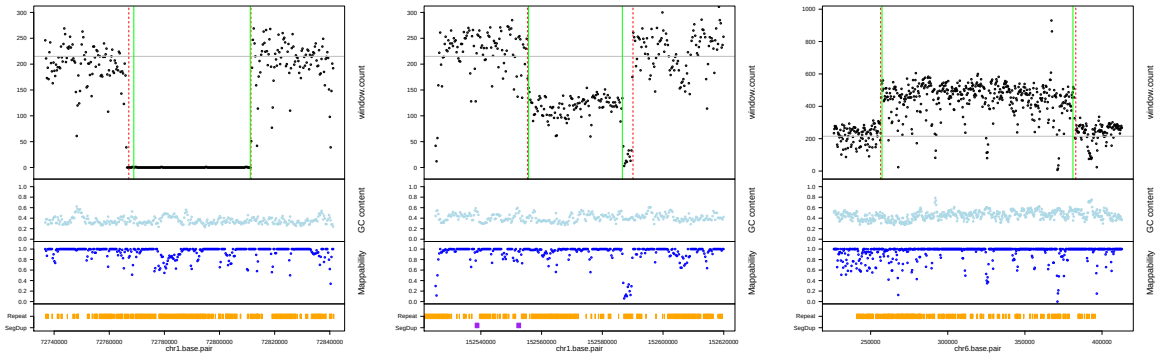


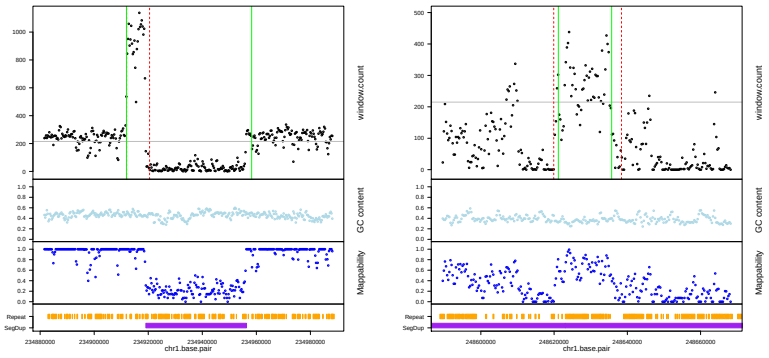
Figure 2.1: GENSENG flowchart

for modeling of copy number. In contrast, existing methods find only two general types of copy numbers, i.e. loss/deletion and gain/duplication. The support for the seven state modeling strategy is evident from examples shown in Figure 2.2, where GENSENG correctly recovered high-confidence CNVs and identified their status as homozygous deletion (Figure 2.2 (a)), heterozygous deletion (Figure 2.2 (b)), and multi-allelic duplications (Figure 2.2 (c)-2.2 (e)). Each subfigure 2.2 (a)-2.2 (e) has 4 panels from top to bottom, and the X-axis of each subfigure indicates genomic position in base pairs. In the first panel, the black dots on the Y-axis indicate read-depth signal; red dashed lines are boundaries from GENSENG prediction; green solid lines are boundaries reported in the high-confidence CNV set (Mills et al., 2011); grey lines are the median read-depth of the chromosome. The GC content and mappability of the region are plotted in the second and the third panels respectively. The fourth panel shows the locations of segmental duplication (purple, from the UCSC hg19 segmental duplication track) and repetitive DNAs (orange, from the UCSC hg19 repeatmask track). Shown here are a homozygous deletion (Figure 2.2 (a)); a heterozygous deletion (Figure 2.2 (b)); a simple and large duplication (Figure 2.2 (c)); a complex duplication (Figure 2.2 (d)) that was predicted to be copy number 6+ and was right-flanked by a large region with a median mappability of 0.2. Finally, (Figure 2.2 (e)) shows a duplication predicted to be copy

number 4 from a noisy region with a median mappability of 0.58, illustrating good sensitivity for detecting duplications by employing simultaneous bias correction and copy number inference.Some discrepancies in the boundaries between the predicted CNVs and the high-confidence CNVs were observed, reflecting technological differences between HTS and microarrays.



(a) a homozygous deletion (b) a heterozygous deletion (c) a four copies duplication



(d) a complex duplication (e) a noise region

Figure 2.2: Example high-confidence CNVs predicted by GENSENG: NA12878 HTS

Second, we used a negative binomial regression model for read-depth and included known confounders as covariates, such that their effects on read-depth were removed. The emission probability of read-depth was made even more robust against read-depth outliers by using a mixture model of negative binomial and uniform distributions, such that any additional biases in the data could be modeled by the negative binomial overdispersion parameter and the uniform distribution. These modeling strategies permit simultaneous bias correction and CNV detection. Some of the benefits of such simultaneous analysis is illustrated in Figure 2.2 (e), which demonstrates good

sensitivity for detecting duplication from a noisy region with a medium mappability score (0.58) after accounting for mappability and additional noises.

Multiple techniques were introduced in GENSENG, including correcting for GC content and mappability, modeling autoregression, fitting a mixture of negative binomial and uniform distributions, and applying quality control to prioritize CNV calls (Methods). To study the effects of these techniques on CNV detection and identify the best-fitting model, we examined the sensitivity and the number of CNV calls made by different partial versions of our method (Table 2.3). We found that correcting for GC content alone was not sufficient; and further correcting for mappability resulted in the most substantial improvement, with gains in both sensitivity and specificity. Quality control including both size and RDA filters substantially improved specificity with minimal loss in sensitivity. The best GENSENG model was selected based on these results and included all aforementioned techniques.

2.2.4 CNV detection method

This triplet of data for each individual genome is input into an integrative hidden Markov model (HMM), which classifies each window to a copy-number state based on maximum a posteriori probability, while simultaneously accounting for sources of bias. The state changes mark the predicted breakpoints of CNVs. Below we present our method and the elements needed in HMM characterization.

Hidden states and transition probability. While microarray analysis suffers from oversaturation at high copy numbers, HTS allows RD-based methods to determine high copy numbers with improved accuracy (Campbell et al., 2008). The state represents the underlying copy number (CN). The state variable $q_t = CN_t$ is hidden and discrete with N possible values, $(0, 1, ..., N-1)$. The total number of hidden states N is implemented as an input parameter of GENSENG and can be freely specified by users. Theoretically, the more high copy number states specified, the more accurate the model becomes. However, a number of practical issues must be considered. For example, specifying more states means longer computing time and, for some datasets, there may not exist sufficient regions from which to estimate parameters. In practice, N

can be derived from the data by K-mean clustering the logarithm of the read-depth.For the HTS datasets used in this study, we assume seven hidden states representing copy numbers of 0, 1, 2, 3, 4, 5, and 6 or more. For homozygous populations such as inbred mice, we assume four hidden states representing copy numbers of 0, 2, 4, and 6 or more. We collapse the duplications with 6 or more copies into one state, because they are difficult to distinguish because of both experimental (reduced signal-to-noise ratio) and computational concerns (having few regions with very high read-depth signal). State transitions proceed from one window to the next according to a first-order time-homogeneous Markov process. The transition probability describes the probability of having a copy-number state change between two adjacent windows. Let π_j be the initial state probability, the probability that the state of the first window is state j. The underlying hidden Markov chain is defined by state transitions $P(q_t|q_{t-1})$ and is represented by a time-independent stochastic transition matrix $A = \{a_{jz}\} = P(q_t = z|q_{t-1} = j)$. Intuitively, the copy number state is unlikely to change for nearby windows but is more likely to change for windows that are far apart.

Emission probability. The hidden copy-number states emit probabilistic outputs at each window, i.e. the observed RD signal representing integer-valued count data. In the absence of sources of bias, sequencing coverage is uniform across the genome such that the emission probability of RD could be modeled by a Poisson distribution with equal mean and variance. In the presence of sources of bias, sequencing coverage is not uniform and the Poisson-distribution assumption fails. To account for biases, the emission probability of RD is modeled as a mixture of uniform distribution and negative binomial (NB), expressed as the following:

$$\begin{aligned} e(t,j) &= P(O_t = o_t|q_t = j) = c/R_m + (1-c)e^{NB}(t,j) \\ &= \frac{c}{R_m} + (1-c)\frac{\Gamma(o_t + 1/(\phi_j))}{o_t!\Gamma(1/\phi_j)}(\frac{1}{1+\phi_j\mu_{tj}})^{1/\phi_j}(\frac{\phi_j\mu_{tj}}{1+\phi_j\mu_{tj}})^{o_t}, \end{aligned}$$

where $e(t,j)$ is the emission probability of a particular observation at a particular time t for state j c is the mixing probability, O_t is a discrete count variable of the observation variable, q_t is

the state variable, o_t is the RD signal for window t, μ_{tj} is the mean RD for window t given state j, ϕ_j is the overdispersion parameter given state j. To describe the negative binomially distributed component, $e^{NB}(t, j)$, we first explain the relationship between the Poisson and the negative binomial distributions. The Poisson distribution imposes that the variance equals to the mean. The negative binomial distribution allows overdispersion. Specifically, if O follows a Poisson distribution with mean μ, and μ follows a gamma distribution, the resulting distribution for O is a negative binomial distribution. The variance of negative binomial distribution is $\mu_t + \phi\mu_t^2$, where $\phi\mu_t^2$ is the overdispersion part of the variance. As $\phi \to 0$, $f_{NB}(o_t; \mu_t, \phi)$ reduces to a Poisson distribution with mean μ_t and variance μ_t. $f_P(o_t; \mu_t) = \frac{\exp(-\mu_t)\mu_t^{o_t}}{o_t!}$. Next, the mean value of the negative binomially distributed component is expressed as a function of a set of covariates to account for confounders.

$$\mu_{tj} \quad = \quad \alpha_0 * (CN_t)^{\beta_1} * (l_t)^{\beta_2} * (g_t)^{\beta_3} \tag{2.1}$$

where t denotes the t^{th} window, j is the index of the copy number state, j emphasizes the dependency of the mean μ_t on the copy number CN_t, l_t is the mappability score, g_t is the GC content. For computational convenience, we set $CN_t = 0.5$ when $j = 0$, and set $CN_t = j$ when $j > 0$.

We then employ a log link function to acknowledge the fact that $\mu_{tj} > 0$ and obtain:

$$\log(\mu_{tj}) \quad = \quad \beta_0 + \beta_1 * \log(CN_t) + \beta_2 * \log(l_t) + \beta_3 * \log(g_t) \tag{2.2}$$

$\beta_0, \beta_1, \beta_2, \beta_3$ are the regression coefficients. Specifically, $\beta_0 = \log(\alpha_0)$, is the intercept parameter and is interpreted as the average level of read-depth signal when all covariates are equal to zero. β_1 is the amount of increase of read-depth for every unit increase of copy number, CN. β_2 is the amount of increase of read-depth for every unit increase of the mappability score, l. β_3 is the amount of increase of read-depth for every unit increase of the GC content, g.

The uniform distribution has a density function $1/R_m$ to model any random fluctuation of read depth, where R_m is treated as a known constant using the largest RD among all windows of the chromosome. When non-overlapping windows are used, the mean RD for each window, μ_{tj}, is modeled by a negative binomial regression model, where the predictors include copy-number state, GC content, and mappability score. A standard HMM assumes the Markov property, $P(q_t|q_{t-1}, q_{t-2}, q_{t-3}, ...q_1) = P(q_t|q_{t-1})$. An additional assumption that is often employed is that the observations are independent given the states, $O_t \perp O_i(i \neq t)|q_t$, which is valid when the windows are non-overlapping. When the windows are overlapping, this assumption is invalid; and instead, the observations are drawn from an autoregression process (Juang and Rabiner, 1985). We have implemented an autoregressive HMM to model this feature of the data. Specifically, a residual term is included as an additional predictor in the negative binomial regression model assuming first order autoregression. Thus in each round of inference, we would first fit the model and obtain the expected read count of state j at window t $\overline{\mu}_{t-1,j}$, and calculate residual $r_{t-1,j} = \log(o_{t-1}) - \log(\overline{\mu}_{t-1,j})$ for $t > 1$ and let $r_{1,j} = 0$. With this extra predictor, we will run GLM again to obtain the true expected read count of state j at window t $\mu_{t,j}$. The additional noise in the data that cannot be explained by variability in GC content and mappability are accommodated by, ϕ_j, the overdispersion parameter of the NB distribution (allowing variance to be larger than mean) and the uniform distribution in the mixture model.

Tuning parameters. Given the HMM topology, the challenge lies in optimizing model parameters given the observed data, a.k.a. HMM training. There are many parameters to be optimized. To reduce computational difficulty, we choose to specify a subset of HMM parameters based on prior knowledge and user preference, including the initial state probability, state transition probability, and the mixing probability in emission probability. These tuning parameters can be influential and should be chosen carefully. The remaining emission parameters, including the coefficients and overdispersion parameters in the NB regression model, are estimated for each dataset.

Parameter estimation. The optimization problem is solved by the Baum-Welch algorithm (Baum et al., 1970), which maximizes the data likelihood for an individual chromosome in iterative steps including initialization, expectation, and maximization. Following Bilmes (Bilmes, 1998), we define the complete-date likelihood and solve the Q function in order to find the maximum likelihood estimates (MLE) of the HMM parameters. In the initialization step, we rely on intuitive guesses as well as empirical values. The initial emission parameters were estimated from the 1000GP and the Mouse Genomes Project datasets where known CNVs are available. These initial emission parameters are saved for the human and mouse genomes respectively and are used for any new sample without prior knowledge of its CNVs. In the maximization step, we obtain maximum-likelihood estimates of emission parameters. Because we fix c and R_m as constant, parameter estimation will only concern the negative binomially distributed component. We apply a weighted negative binomial regression model, where the weights are posterior probabilities for each window belonging to a particular copy-number state, given the observed data of an entire chromosome. These weights represent current knowledge of the probabilistic classification of a window to copy-number state and are updated in the expectation step. While included as a predictor in the regression model, the copy number is the hidden variable to be inferred from the observed data. Intuitively, by using posterior probability as regression weights, we are able to partition the observed RD across all hidden states, proportional to the likelihood. The weighted NB regression model is fitted by alternately estimating regression coefficients using iteratively reweighted least squares and estimating the overdispersion parameter using a Newton-Raphson method. In the expectation step, we update the forward, backward, and posterior probability given the current estimates from the maximization step. The expectation and maximization steps iterate until the convergence criterion (smaller than 10^{-6} change in the log-likelihood) is reached.

CNV calling. Using the parameters at convergence, first we obtain the final estimates of the posterior probability for each window belonging to a particular state, given the observed data from the entire chromosome. Second, we assign the final estimate of copy number for each window using the state with the largest posterior probability. The state changes mark the predicted

breakpoints of CNVs. The confidence score of a CNV region is computed as the sum of the posterior probabilities of all windows enclosed within the breakpoints. Next, a two-step merging algorithm is carried out to refine the boundaries of the CNVs.

Prioritization of CNV calls. A CNV quality control step can be applied to remove CNVs predicted with the lowest confidence. We recommend removing predicted CNVs shorter than 800bps (i.e. removing those that appear in only one window as shown in Table 2.3), or predicted CNVs with an average mappability lower than 0.3 (i.e. removing those that cannot be confidently predicted as shown in Figure 2.6 (b)). An additional prioritization approach was implemented via the read-depth-accessibility (RDA) statistic, which reflects the signal-to-noise ratio of a predicted CNV region after accounting for known confounders in read-depth. The term of read-depth-accessibility was first coined by Abyzov et al. (Abyzov et al., 2011), but for a different purpose. The RDA statistic is computed in 3 steps: (1) after CNV calling, identify all compatible copy-number-neutral windows whose GC-content and mappability scores are the same as those from the region of interest; (2) calculate the average window counts from (1) as the expected read-depth for the region of interest; (3) obtain the RDA by dividing the observed read-depth by the expected read-depth for the region of interest. Using a copy number of two as normalization for copy-number-neutral autosomal regions, the theoretical signal-to-noise ratios are 0, 0.5, 1.5, and 2 for copy numbers of 0, 1, 3, and 4, respectively. Therefore, a region is considered to be read-depth-accessible if its RDA value is lower than 0.5 for homozygous deletions, lower than 0.75 for heterozygous deletions, and greater than 1.25 for duplications. In general, we recommend removing CNVs predicted from regions that are not read-depth-accessible (e.g. if its RDA values range between 0.75 to 1.25). In addition, we recommend ranking the predicted regions by their RDAs, where a higher signal-to-noise ratio reflects higher confidence that the predicted CNVs are correct; this is analogous to ranking by fold-change in gene-expression analysis.

2.2.5 Performance assessment

While the high-confidence dataset compiled by Mills et al. (Mills et al., 2011) indicates where the true positive CNVs are for HapMap individuals, the true negatives are unknown.

Therefore, we used two approaches to assess our methods performance. First we conducted a simulation to estimate the sensitivity and specificity to detect CNVs. Then we analyzed the high-coverage 1000GP trio data, where we estimated the sensitivity using high-confidence CNVs and used the total number of base pairs or calls as a surrogate measure for specificity. For comparison, we applied CNVnator (Abyzov et al., 2011) in parallel, using its recommended parameter setup and QC filter. The methodology differences between GENSENG and CNVnator are detailed in Table 3.2. The main differences are in bias correction and segmentation techniques.

Simulation. We simulated two datasets for performance assessment. The first simulation directly generated read-depth data (Yoon et al., 2009). Using chromosome 1 from NA12878 as a template, we implanted 76 high-confidence CNVs (25 duplications and 51 deletions) (1) by assigning a copy number of four to any window that overlapped the duplications and a copy number of zero to any window that overlapped the deletions. All other windows were assigned to have a copy number of two. The covariate matrix (the assigned copy number, mappability score, and GC content of each sliding window) and coefficient vector were passed to the garsim function from R/gsarima to simulate the read-depth for each window. The garsim model we applied was the negative binomial distribution with the log link function, where the autoregressive parameter was set to 0.6, the zero correction parameter was set to zq1, and the inverse of the overdispersion parameter was set to 0.01.

The second simulation mimicked a sequencing experiment to generate paired-end reads from a CNV containing a hypothetical chromosome. To simulate reads, we used chromosome 1 of the reference human genome as a template and modified the template sequence based on the 76 high-confidence CNVs (51 deletions and 25 duplications) (Mills et al., 2011). For any deletion, we removed the corresponding sequence of the deleted DNA, and for any duplication, we inserted an extra copy of the duplicated sequence. As a result, the implanted deletions were copy number 0 deletions, and the implanted duplications were copy number 4 duplications. Among the 76 high-confidence CNVs, 8 deletions and 4 duplications overlap with other deletions or duplications. Thus a total of 64 independent CNVs were implanted (43 deletions and 21 duplications) into the

chromosome 1. Second, after the CNV-containing hypothetical chromosome was created, we applied the sequencing simulator, wgsim, as implemented in SAMTools (Li et al., 2009) to generate 36bp paired-end short reads. For wgsim simulation, the mean value of the outer distance between the two ends was set to 200, the standard deviation was set to 20, and the sequencing error model was the empirical error model of the Illumina sequencing platform. A total of 150 million paired-end reads were generated, which gave an average sequencing coverage of 40x. Third, we used BWA (Li and Durbin, 2009) to map the reads to the unmodified reference human genome. The resulting alignment file was use as input to apply GENSENG (Szatkiewicz et al., 2013) and CNVnator (Abyzov et al., 2011). Among CNVs predicted by either approach, a true discovery was defined when a predicted CNV overlapped with at least 50% of a simulated CNV and had the same copy number.

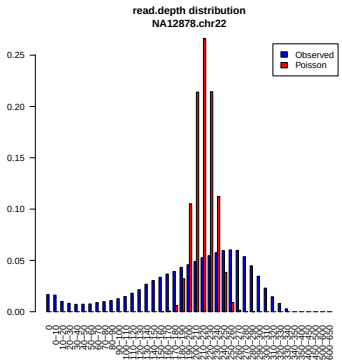
1000GP data. We analyzed the high-coverage sequencing data for the CEU trio from the 1000GP. To facilitate the comparison between the predicted CNVs and the high-confidence CNVs, which only provide deletion and amplification calls rather than the particular copy number, we defined deletions as any GENSENG calls where the inferred copy numbers were 0 or 1, and duplications as any calls where the inferred copy numbers were greater than 2. Sensitivity was calculated by dividing the number of total base pairs of the overlapping events ($>$1bp overlap, or $>$50% reciprocal overlap with the high-confidence CNVs) by the total number of high-confidence CNVs.

Performance on low coverage data: The native coverage of both our simulated data and the 1000GP high-coverage data is $\sim$40x. To identify the lower bound that GENSENG can handle, we applied GENSENG to data with varying sequencing coverage and compared the performance to that based on the native coverage using the same evaluation metrics. First, we repeated the simulation process as described earlier with the targeted coverage been set as 5x, 10x, 20x, 30x, and 40x. To test the consistency of our simulation, we also simulated data at 40x coverage for 10 times and observed replicable results (data not shown). Second, using the DownsampleSam.jar tool from

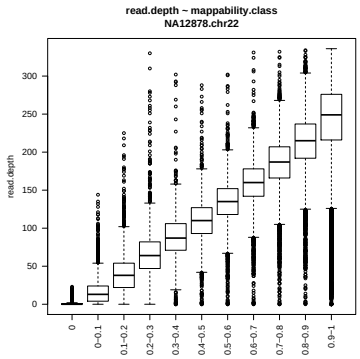
Picard, we down-sampled the high coverage 1000GP data from NA12878 and achieved a series of sequencing coverage of 5x, 10x, 20x, 30x, 40x.

2.3 Results

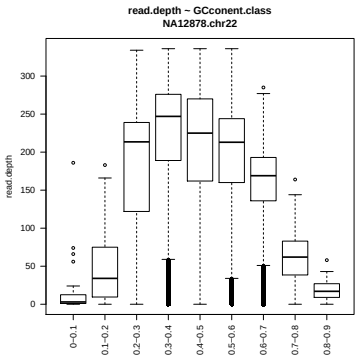
2.3.1 Evaluation of experimental biases in HTS



(a) The observed distribution has larger variance than Poisson distribution

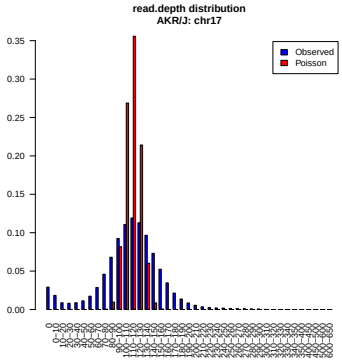


(b) A positive correlation is observed between read count and mappaiblity

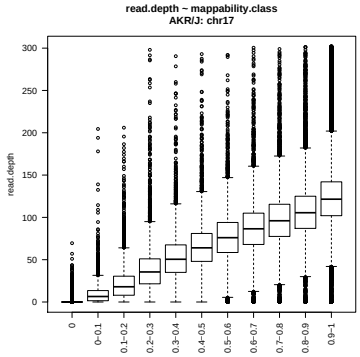


(c) A non-linear correlation is observed between read count and GC content

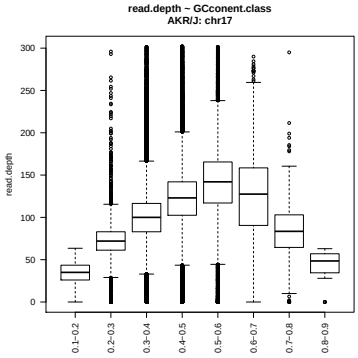
Figure 2.3: Relationship between read-depth and known confounders: NA12878



(a) The observed distribution has larger variance than Poisson distribution



(b) A positive correlation is observed between read count and mappaiblity



(c) A non-linear correlation is observed between read count and GC content

Figure 2.4: Relationship between read-depth and known confounders: AKR/J

We demonstrated the experimental biases observed in multiple HTS samples in the following paragraph. We studied the relationship between read count (the number of fragments in each genomic window, a.k.a window count) and biases (GC content and mappability) from

chromosome 22 of human sample NA12878 and chromosome 17 of mouse inbred strain AKR/J. Under idealized scenarios, HTS read-depth is expected to follow a Poisson distribution with variance equal to the mean. However, we found that the observed variance is much greater than the mean (Figures 2.3 (a) and 2.4 (a), indicating substantial deviation from the Poisson distribution. We found that some of the non-uniformity in read-depth is caused by genome-wide variability in mappability and GC content. Figures 2.3 (b) and 2.4 (b) show a positive correlation between mappability and read-depth, where low mappability scores indicate a higher proportion of repetitive sequences, resulting in lower read-depth; high mappability scores indicate a higher proportion of unique sequences, resulting in higher read-depth. Figures 2.3 (c) and 2.4 (c) show a non-linear relationship between GC content and read-depth, where sequences with extreme GC content (low or high) tend to have lower read-depth. In addition to the general trends observed across various datasets, we found that the curves of read-depth vs. GC content varied from sample to sample. For example the peak of the mouse sample slightly shifted to the right (Figure 2.4 (c)). The fractions of mappable bases in the genomes were found to be 80%-90% and increased moderately as the read-length increased (Table 2.1). Lastly, we found that the non-uniformity in read-depth could not be explained solely by GC content or mappability. We examined the read-depth distribution from compatible windows that had the same GC content, the same mappability scores, and were mostly likely copy-number-normal (e.g. did not overlap any high-confidence CNVs or other candidate CNVs). Then we compared the observed distribution from these compatible windows to the theoretical expectations. If GC content and mappability were the only sources of biases, the observed distribution should have closely followed a Poisson distribution. However, Figure 2.5 suggests that the Poisson distribution still fails (the red curve is far apart from the blue curve), because it restricts variance to equal the mean; on the other hand, the negative binomial distribution fits the data well (the green curve is close to the blue curve), because its overdispersion parameter accommodates additional sources of noise in the data.

[1] see Section 2.2.1 for reference genomes. To generate the results in this table, sequences from chrN_random and chrUn were excluded.

Table 2.1: Fraction of mappable bases in human and mouse genomes

Organism [1] (Gb)	Read Length (nt) used to compute mappable base	Genome Size[2]	Mappable Sequence Size (Gb)[1]	Mappable Sequence Percentage[1]
Mus musculus	2.655	54	2.284	86.0%
Mus musculus	2.655	100	2.389	90.0%
Homo sapiens	3.095	36	2.490	80.4%
Homo sapiens	3.095	100	2.765	89.3%

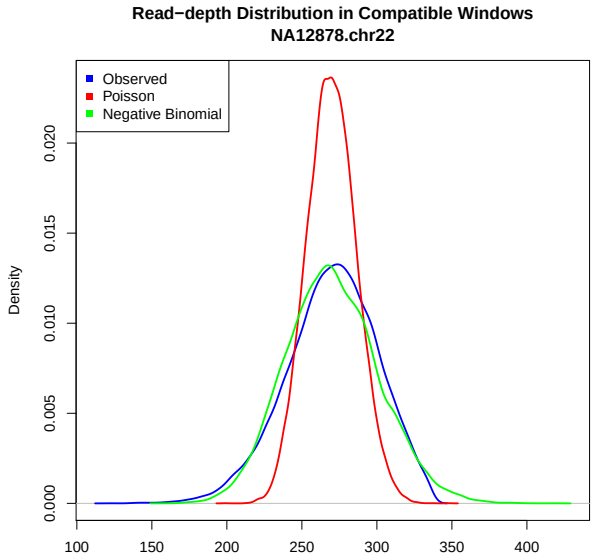
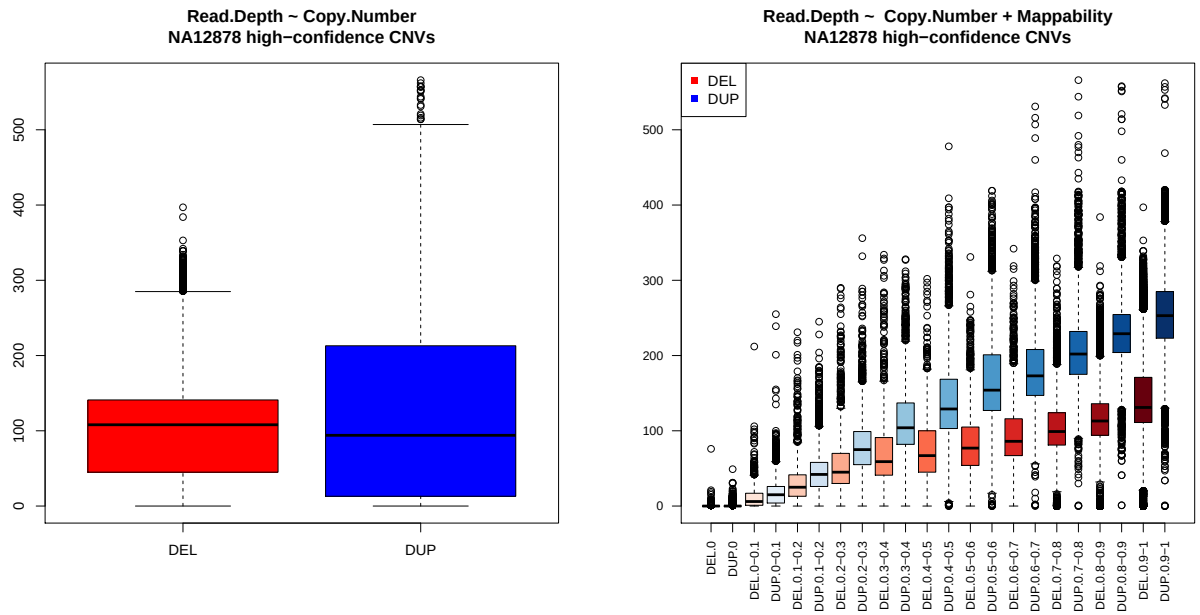


Figure 2.5: Read-depth distribution after accounting for known confounders

To investigate how experimental biases impact copy-number inference, we examined the relationship between mappability and read-depth in high-confidence CNVs (610 deletions and 261 duplications) (Mills et al., 2011). The relationship was visualized in Figure 2.6. Figure 2.6 (a) shows the boxplot of read-depth from windows mapped to the 610 high-confidence deletions (red) and 261 high-confidence duplications (blue), suggesting a similar read-depth distribution between deletions and duplications and no power in detecting CNVs. Figure 2.6 (b) shows the boxplot of read-depth stratified by mappability classes, color-coded such that darker shades reflect higher mappability. The labels of the X-axis indicate the CNV class (DEL: deletions; DUP: duplications) and mappability class. For example, label (DEL.0.2-0.3) indicates windows from the

[2]see Section 2.2.2 for our algorithm used to compute mappable base. The fractions reported here are similar to other studies in the literature. For example, Rozowsky et al (Rozowsky et al., 2009) used an indexing algorithm and found the fractions of mappable bases based on 30bp read is 81% in the mouse genome and 79.6% in the human genome (Table 1 of Rozowsky et al (Rozowsky et al., 2009)).



(a) Without stratified mappability

(b) With stratified mappability

Figure 2.6: Relationship between read-depth and mappability in high-confidence CNVs.

high-confidence deletions and with mappability score ranging from 0.2 to 0.3. Within each mappability class, duplications show higher mean read-depth than deletions, suggesting that correction for mappability improves the ability to detect CNVs. Furthermore, when mappability falls below 0.3, read-depth distribution becomes increasingly similar between deletions and duplications, suggesting that the ability to detect CNVs in those regions is limited. For example, for windows with mappability ranging from 0.2 to 0.3, approximately 50% of windows in the duplication regions had read-depths equal to or lower than the average read-depth from compatible copy-normal regions; approximately 20% of windows in the deletion regions had read-depths equal to or higher than the average read-depth from compatible copy-normal regions.From the picture we could conclude that without accounting for mappability, duplicated and deleted regions could not be distinguished, because the read-depth distributions were very similar (Figure 2.6 (a)). However, after stratification by mappability scores, the mean read-depths became significantly different between duplicated and deleted regions within a mappability class, such that these CNVs could be recovered (Figure 2.6 (b)). This observation suggests that it is important to jointly estimate copy number and the effect of confounding factors. Copy number inference made without correcting for

bias may lead to systemic errors. Furthermore, as shown in Figure 2.6 (b), when mappability scores are extremely low (e.g. $<$0.3), too few reads can be confidently aligned to those regions, and consequently CNVs cannot be confidently predicted. This observation suggests that a CNV quality-control filter based on mappability score could be applied to reduce false positive predictions.

2.3.2 Performance assessment and comparison

To benchmark the performance of GENSENG, we carried out two sets of simulations. In the first simulation, we applied GENSENG to simulated read-depth data (sequence-fragment counts across tiled windows) for a single chromosome. As illustrated in Figure 2.7, the sensitivity was computed as the total true CNVs detected divided by the total true CNVs. In this study, the criterion for detected were either more than 1bp in the true CNV was overlapped by predicted CNV or more than 50% bps of true CNV were overlapped by predicted CNV (default $>$ 50% overlapped unless explicitly noted). FDR was computed as the total falsely predicted CNV calls divided by the total predicted CNV calls. Similarly there were two criterion for false prediction: $>$ 1bp and $>$ 50% overlaped with any true CNVs (again default $>$ 50% overlapped unless explicitly noted). For 76 simulated CNVs, we observed 88% sensitivity and 100% specificity. The remaining 12% simulated CNVs (3 duplications and 6 deletions) did not pass CNV quality-control (QC) filters (read-depth-accessibility (RDA) filter and mappability$<$0.3).

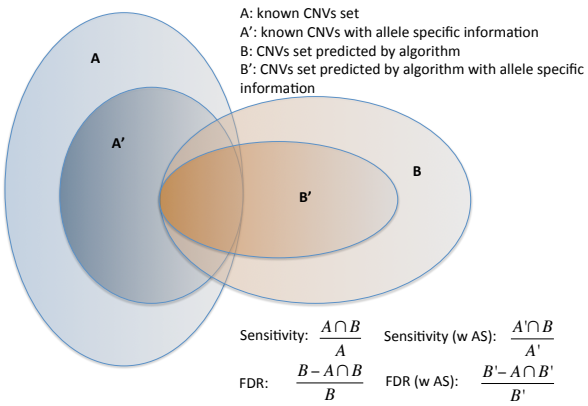


Figure 2.7: Illustration of sensitivity and FDR calculation

In the second simulation, we applied both GENSENG and CNVnator to the simulated sequence reads (instead of window read counts) for a single chromosome; and we compared the sensitivity and specificity between GENSENG and CNVnator (Table 2.2). Before any CNV-QC-filter was applied, GENSENG produced 1478 fewer false CNV calls (69% fewer false deletions and 56% fewer false duplications), demonstrating a better specificity. A total of 12 simulated CNVs (10 deletions and 2 duplications) were not detected by GENSENG; and they were missed because they were smaller than the minimum size of CNVs detectable by GENSENG (i.e. $<$800bp), or had mappability < 0.3 (i.e. unreliable regions). After the recommended CNV-QC-filters were applied (read-depth-accessibility (RDA) filter for GENSENG and the default q0 filter for CNVnator), GENSENG outperforms CNVnator in both sensitivity and specificity. Specifically, for 43 simulated deletions, GENSENG had 77% sensitivity, 7% higher than CNVnator. For 21 simulated duplications, GENSENG had 90% sensitivity, 33% higher than CNVnator. The specificity for duplications was 100% for both GENSENG and CNVnator. Both methods made false deletion discoveries, but GENSENG had better specificity (1328 fewer false deletions, 12% lower FDR than CNVnators). Increasing the stringency of the CNV-QC-filter by removing CNVs with mappability < 0.3 further improved GENESENGS specificity (9% FDR for deletions, or 43% lower FDR than CNVnators based on the same stringent CNV-QC-filter), while maintaining its good sensitivity (67% for both deletions and duplications, or 15% higher sensitivity than CNVnators). This result and the result from Figure 2.6(B) suggest the usefulness of the mapability filter. We also found that the RDA statistic (i.e., signal-to-noise ratio after accounting for confounders) computed for each CNV is an effective approach to correctly prioritize the prediction made by GENSENG or CNVnator. Most remaining false positive deletions in the CNVnator dataset were likely influenced by sources of bias and the 52% FDR is a reasonable estimate for CNVnator based on the experimental validation conducted by the 1000GP (reported to range from 14.3% to 74.1%) (Mills et al., 2011; Abecasis et al., 2012). In summary, the simulation studies demonstrate that GENSENG outperforms CNVnator, suggesting that integrating bias correction from multiple sources and copy-number inference is a desired strategy for read-depth-based CNV detection.

Table 2.2: Performance assessment based on simulated sequencing data for a chromosome

		Deletions				Duplications			
Detection Method	Post-detection CNV-filter	# Simulated TRUE CNV	# Predicted CNV calls	Sensitivity[3] (# TRUE CNV detected)	FDR[4] (Number of false prediction)	# Simulated TRUE CNV	# Predicted CNV calls	Sensitivity[3] (# TRUE CNV detected)	FDR[4] (# False prediction)
CNVnator	None	43	2171	0.94(40)	0.98(2130)	21	40	1.00(21)	0.23(9)
GENSENG	None	43	690	0.77(33)	0.95(657)	21	26	0.90(19)	0.12(4)
CNVnator	q_0-filter[5]	43	1560	0.70(30)	0.98(1530)	21	19	0.57(12)	0(0)
CNVnator	q_0[5] +(RDA+map)[7]	43	23	0.53(23)	0.52(25)	21	18	0.52(11)	0(0)
GENSENG	RDA[6]	43	235	0.77(33)	0.86(202)	21	22	0.90(19)	0(0)
GENSENG	(RDA+map)[7]	43	32	0.67(29)	0.09(3)	21	16	0.67(14)	0(0)

We incorporated various techniques for bias correction, i.e. to account for GC content and mappability, to model the auto-regression in the read-depth data computed from overlapping windows, to account for additional noises by fitting a mixture of negative binomial and uniform distributions, and to apply quality control to prioritize the CNV calls. As these techniques incorporated incrementally, there are five partial versions of our algorithm as shown in Table 2.3. In the first step the partial algorithm only corrects for GC content. In the next step, the partial algorithm corrects for GC content and mappability together. Similarly, the auto-regression and the mixture modeling are added incrementally. The algorithm becomes our full algorithm after the CNV quality control step is applied where we (1) removed predicted CNVs that are shorter than 800bps (i.e. those that appear in only one window), and (2) additionally removed predicted CNVs with RDA value ranging between 0.75-1.25. To evaluate the effect of different techniques on CNV inference, we used the high confidence CNVs reported in Mills et al (Mills et al., 2011) (610 deletions and 462 duplications) to examine the performance of different partial versions of our algorithm. Sensitivity is computed as the total number of high-confidence CNVs that were overlapped by the GENSENG predicted calls (>50% overlapping) divided by the total number of high- confidence CNVs. We then used the total number and total base pairs of the GENSENG

[3]Note that a single true duplication could be overlapped by >1 predicted calls.

[4]Also a single predicted calls may also be overlapped by >1 true call

[5]the default q0 filters removes any predicted calls that have >50% reads with zero-valued MAPQ (i.e. reads with multiple mapping locations).

[6]Our RDA filter removes any predicted calls that have RDA values ranging between 0.5-1.25.

[7]Our stringent filter removes any predicted calls that have RDA values ranging between 0.5-1.25 or mappablility < 0.3. As all simulated deletions were homozygous deletions, 0.5 was used as the lower threshold of RDA.

Table 2.3: Evaluation of the effects of modeling techniques using NA12878

Model		Deletions			Duplications		
Step	Technique	# Calls	Mbps spanned	Sensitivity	# Calls	Mbps spanned	Sensitivity
1	+ GC Content	32402	215.8	0.59(362/610)	1350	65.7	0.12(32/261)
2	+ Mappability	11949	62.5	0.72(438/610)	8471	94.3	0.23(60/261)
3	+ Auto Regression	7815	60.8	0.71(434/610)	3051	79.8	0.20(53/261)
4	+ Random uniform distribution	9364	62.7	0.73(447/610)	3529	80.8	0.20(53/261)
5(1)	+ CNV quality control: size filter	7734	61.8	0.73(447/610)	2009	80.1	0.19(50/261)
5(2)	+ CNV quality control: RDA filter	5370	48.8	0.73(446/610)	1577	68.1	0.17(45/261)

predicted calls as a surrogate measure of specificity. As shown in Table 2.3, correcting for GC content alone resulted in low sensitivity and specificity. After various techniques were added step-wisely, both the sensitivity and the specificity were improved. Correcting for mappability resulted in the most substantial improvement in both sensitivity and specificity. Quality control of the predicted CNVs substantially improved specificity with minimal loss in sensitivity. The full algorithm demonstrated the best performance and thus was used as the release of our GENSENG software.

To further evaluate GENSENGs performance, we analyzed the 1000GP data. First, we applied both GENSENG and CNVnator to the high-coverage HTS data for NA12878 and focused on calling autosomal CNVs (Table 2.4). Sensitivity was estimated by comparing the predicted CNVs to the high-confidence CNVs from Mills et al. (Mills et al., 2011) (610 deletions and 261 duplications). GENSENG gave an overall sensitivity of 56% (73% for deletions and 17% for duplications) using the 50% reciprocal overlap criterion. In contrast, CNVnator gave a lower overall sensitivity of 50% (64% for deletions and 16% for duplications) using the same criterion. Approximately 87% of the high-confidence CNVs were obtained from high-density microarrays based on changes in probe intensities (Conrad et al., 2010; McCarroll et al., 2008). For these CNV regions, the evidence for changes in read-depth may or may not be observed from HTS data (Abyzov et al., 2011) (also see Section 2.4), which could be predicted by the read-depth-accessible (RDA) statistic. Similarly to Abyzov et al (Abyzov et al., 2011), we found that $\sim$76% (462 out of 610) high-confidence deletions were read-depth accessible (i.e. RDA$<$0.75) from the HTS data; whereas only $\sim$21% high-confidence duplications (53 out of 261) were read-depth accessible (i.e. RDA$>$1.25). Given these observations, we then recomputed sensitivity by comparing the predicted CNVs to the high-confidence CNVs that are read-depth accessible (462 deletions and 53

Table 2.4: Performance assessment based on NA12878 HTS data

		Deletions			Duplications		
Method	Post-detection CNV-filter	# High-confidence CNVs[8]	# Predicted calls (Mbps spanned)	Sensitivity[9] (# High-confidence CNV detected)	# High-confidence CNVs[8]	# Predicted calls (Mbps spanned)	Sensitivity[9] (# High-confidence CNV detected)
CNVnator	q0 filter[10]	610	4105(142.1)	0.64(393)	261	788(9.5)	0.16(42)
GENSENG	RDA[11]	610	5370(48.8)	0.73(446)	261	1577(68.1)	0.17(45)
GENSENG	(RDA+map)[12]	610	4087(11.7)	0.7(427)	261	984(38.5)	0.15(39)

duplications), which yielded an overall sensitivity of 90% for GENSENG and an overall sensitivity for 79% for CNVnator. (The 79% sensitivity is similar to that reported by the authors of CNVnator (Abyzov et al., 2011)). The high-confidence CNV set (Mills et al., 2011) does not provide information on the true negatives needed to assess specificity; thus we focused on calibrating sensitivity as described above and used the volume (i.e. the total number and total base pairs) of the predicted CNVs as a surrogate measure of specificity. We found that the predicted volumes are comparable between the two methods.

Then, we applied GENSENG and CNVnator to the 1000GP HTS data from the CEU trio (NA12878, NA12891, NA12892); and we evaluated the sensitivity for detecting deletions as compared to the deletions from Handsaker et al. (Handsaker et al., 2011), which represent the combined CNV calls from 1000GP (Mills et al., 2011). We applied the default filters. CNVnator filter: the default q0 filters removes any predicted calls that have >50% reads with zero-valued MAPQ (i.e. reads with multiple mapping locations). GENSENG filter removes any predicted calls that have RDA values lower than 0.75. Sensitivity was calculated by dividing the number of the overlapping events (>1bp or >50% overlap with the high-confidence CNVs) by the total number of high-confidence CNVs. We found that an average of 49% deletion calls from Handsaker et al. (Handsaker et al., 2011) intersected with GENSENG calls (Table 2.5). In contrast, we found an

[8]It does not provide information on the true negatives needed to assess specificity. Thus we focused on calibrating sensitivity and used the total number and total base pairs of the predicted CNV calls as surrogate measure of specificity.

[9]>50% overlapping

[10]the default q0 filters removes any predicted calls that have >50% reads with zero-valued MAPQ (i.e. reads with multiple mapping locations).

[11]GENSENG filter removes any predicted calls that have RDA values ranging between 0.75-1.25.

[12]Stringent GENSENG filter removes any predicted calls that have RDA values ranging between 0.75-1.25 or mappablility < 0.3. As deletions were either homozygous or heterozygous, 0.75 was used as the lower threshold of RDA.

Table 2.5: Sensitivity to detect deletions in CEU trio data

Genome	# High Confidence Deletion[13]	# Deletions calls (Mbps spanned)		Sensitivity>1bp overlap		Sensitivity >50% overlap	
		GENSENG	CNVnator	GENSENG	CNVnator	GENSENG	CNVnator
NA12878	2301	5370(48.8)	4105(142.1)	0.49	0.39	0.49	0.38
NA12891	2200	4765(88.1)	2656(131.3)	0.50	0.38	0.50	0.37
NA12892	2055	4295(45.0)	2268(128.0)	0.49	0.34	0.49	0.34
Average	-	-	-	0.49	0.37	0.49	0.36

average of ∼37% intersected with CNVnator calls. The deletions reported in Handsaker et al. (Handsaker et al., 2011) were derived from the results from the 19 algorithms used by the 1000GP (Mills et al., 2011) and contained many deletions that are smaller than the minimum size of CNVs (<800bp) detectable by GENSENG. For NA12878, 89% of deletions (1026 out of 1148) from Handsaker et al. (Handsaker et al., 2011) missed by GENSENG were due to the size. Similarly, for NA12891, 51% of deletions (547 out of 1072) were missed due to the size, and for NA12892, 51% of deletions (523 out of 1030) were missed due to the size. In summary, the analyses of the 1000GP data confirm that GENSENG have better detection sensitivity than CNVnator and suggest similar or better specificity.

To identify the lower bound that GENSENG can handle, we applied GENSENG to data with varying sequencing coverage generated by simulation and compared the performance to that based on the native coverage (40x) using the same evaluation metrics (as was done in 2.2). We repeated the simulation process as described earlier with the targeted coverage been set as 5x, 10x, 20x, 30x, and 40x. As in Table **??**, stringent GENSENG filter removes any predicted calls that span only one window, have RDA values ranging between 0.5-1.25 or mappablility ¡ 0.3. The results in Table 2.6 demonstrated that: (1) higher sequencing coverage improves CNV detection power; (2) the lower bound of sequencing coverage that yield reasonably good performance of GENSENG is 10x (70% sensitivity and 6% FDR for deletions; 57% sensitivity and 0 FDR for duplications); (3) GENSENG could potentially work on data sequenced to as low as 5x but with much reduced sensitivity (fell by 14% for deletions and fell by 10% for duplications). We also applied GENSENG to datasets down-sampled from 1000GP and compared the performance to that based on the native coverage (40x) using the same evaluation metrics (as was done in Table 2.4). We used the

[13]It was generated from (Mills et al., 2011; Handsaker et al., 2011). The coordinates of reported deletions were translated from NCBI36 to NCBI37 using liftOver.

Table 2.6: Performance on datasets with varying sequence coverage: simulation

Coverage	Deletions				Duplications			
	# Simulated TRUE CNV	# Predicted CNV	Sensitivity	FDR	# Simulated TRUE CNV	# Predicted CNV	Sensitivity	FDR
5x	43	25	0.58	0	21	11	0.52	0
10x	43	32	0.70	0.06	21	13	0.57	0
20x	43	34	0.72	0.09	21	15	0.57	0
30x	43	35	0.72	0.11	21	15	0.57	0
40x	43	35	0.72	0.11	21	17	0.62	0

Table 2.7: Performance on datasets with varying sequence coverage: simulation

Coverage	Deletions			Duplications		
	# High Confidence CNVs	# Predicted CNVs (Mbps spanned)	Sensitivity	# High Confidence CNVs	# Predicted CNVs (Mbps spanned)	Sensitivity
5x	610	745 (5.42)	0.38	261	482 (21.13)	0.12
10x	610	1344 (6.51)	0.53	261	769 (47.16)	0.14
20x	610	1939 (6.76)	0.57	261	941 (55.76)	0.14
30x	610	2488 (7.64)	0.62	261	1042 (30.55)	0.14
40x	610	5071 (11.73)	0.70	261	984 (38.58)	0.15

DownsampleSam.jar tool from Picard to down-sample the high coverage 1000GP data from NA12878 and achieved a series of sequencing coverage of 5x, 10x, 20x, 30x, 40x. As in Table 2.4, stringent GENSENG filter removes any predicted calls that span only one window, have RDA values ranging between 0.75-1.25 or mappablility < 0.3. The results in Table 2.7 demonstrated that (1) higher sequencing coverage improves CNV detection power; (2) the lower bound of sequencing coverage that yield reasonably good performance of GENSENG is 10x (53% sensitivity for deletions; 14% sensitivity for duplications); (3) GENSENG could potentially work on data sequenced to as low as 5x but with much reduced sensitivity (fell by 32% for deletions and fell by 3% for duplications). In short, we demonstrated that, as expected, higher sequencing coverage improves CNV detection power; and that the lower bound of sequencing coverage that yields reasonably good performance of GENSENG is 10x (Tables 2.6 and 2.7). GENSENG could potentially work on data sequenced to as low as 5x but with much reduced sensitivity (fell by 33% for deletions and fell by 10% for duplications, Tables 2.6 and 2.7).

2.3.3 Application to other HTS data

As a proof of concept, we applied GENSENG to whole-genome HTS data from human and mouse samples and evaluated the validity of its prediction using an allele-sharing principle as well as additional genetic information available in our data. By allele-sharing, we mean the following. From a sequencing study, genetic mutations can be readily detected with a broad spectrum of allele frequency, ranging from singleton variants that are unique to individual genomes to variants

observed in multiple genomes. A variant shared among multiple genomes could arise from the inheritance of the same ancestral allele, i.e. identity by descent (IBD), such that shared variants could receive higher detection confidence. The idea of searching for shared variation to increase the power of CNV detection was previously explored by Handsaker et al. (Handsaker et al., 2011) in the 1000GP samples from low-coverage population-scale sequencing. In our study, we first predicted CNVs from individual genomes and then identified shared CNVs that could arise from IBD as an evaluation of GENSENGs performance.

The three human individuals affected by bipolar disorder that we examined were cousins, and therefore they were expected to share approximately 1.5% of their genomes (1.5% IBD). If a genomic region is IBD, we expect to see a similar read-depth pattern for that region in each individual genome and in the pooled reads from all individuals. If the IBD region contains a true CNV, this CNV could be detected based on higher or lower than expected read-depth using either individual alignment files or the pooled alignment file. Known confounders such as genomic GC content and mappability could also create similarity in read-depth pattern across different genomes and consequently predict CNVs that are shared among them. However, since GENSENG accounts for these confounding factors while inferring copy-number states, shared CNVs that arise from such artifacts have been minimized. In summary, GENSENG identified a total of 831 candidate CNVs that are shared among the three cousins. Shared CNVs, especially those unique to this pedigree, could indicate an enrichment of high-risk disease alleles, and these CNVs are reported elsewhere. To illustrate the utility of GENSENG, we showed two examples of deletions detected from human HTS datasets (Figure 2.8,2.9). The legend of Figure 2.8 is as follows. There are 6 panels from the top to the bottom. The X-axis of each panel indicates genomic position in base pair but the specific coordinates are not shown. The first panel: read-depth computed from individual spa5w; the second panel: read-depth from individual spa42w; the third panel: read-depth from individual spa105w; the fourth panel: read-depth computed after pooling all the reads together from spa5w, spa42w, and spa105w. The GC content and mappability of the region are plotted in the fifth and the sixth panels respectively. In the first through the fourth panels, the black dots in Y-axis

indicate read-depth signal; the red dashed lines are the boundaries from GENSENG prediction; the grey lines are the median read-depth of the chromosome. The legend of Figure 2.9 is as follows. There are 5 panels from the top to the bottom similarly to Figure 2.8. The first panel: read- depth from individual spa5w; the second panel: read-depth from individual spa42w; the third panel: read-depth from individual spa105w; the fourth panel: GC content; the fifth panel: mappability. In individual spa5w (first panel), the steel blue dots indicate read-depth signal from a predicted duplication whose boundary is enclosed by the red dashed lines. Figure 2.8 shows an example of shared deletion, also included in the 1000GP SV release (Mills et al., 2011; Handsaker et al., 2011). This suggests that alleles segregated in the general population at an appreciable frequency (e.g. $>1\%$ in 1000GP samples) would generally be shared among multiple individuals sequenced (Handsaker et al., 2011). In contrast, Figure 2.9 shows an example of a singleton duplication that warrants further experimental validation.

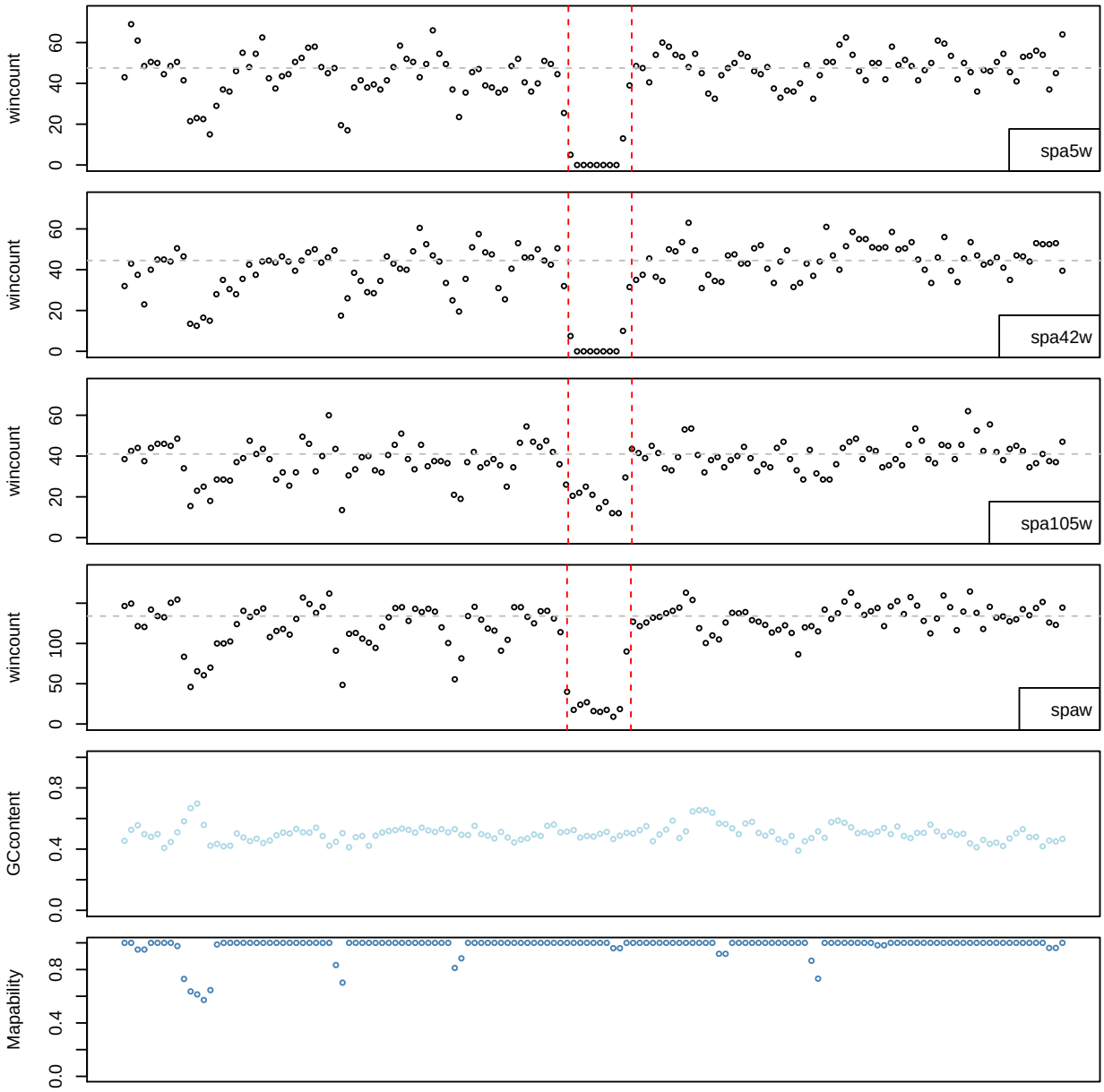


Figure 2.8: Example shared deletion identified from the human HTS dataset.

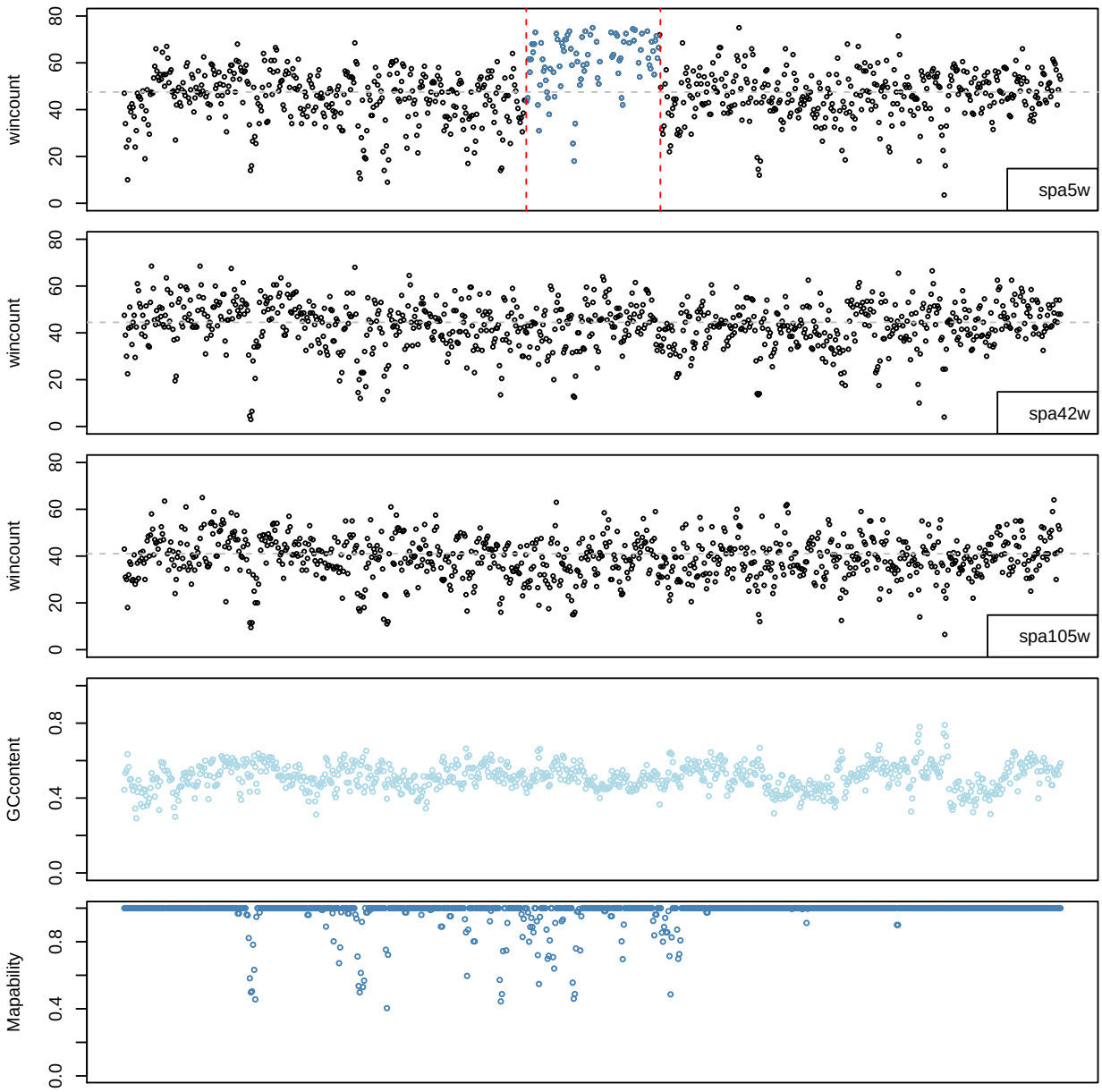


Figure 2.9: Example private deletion identified from the human HTS dataset.

Similarly, we examined shared CNVs in the mouse genome. A genome-wide haplotype and IBD map has been established in 100 classical mouse strains using high-density SNP genotypes (Yang et al., 2011). Strains belonging to the same haplotype in a genomic region had greater than 99% sequence identity and were considered IBD over that interval (Yang et al., 2011). To illustrate the utility of GENSENG, we showed two examples of shared CNVs that stem from IBD detected from mouse HTS datasets (Figure 2.10,2.11). The legend of Figure 2.10 is as follows. There are 6

panels from the top to the bottom. The X-axis of each panel indicates genomic position in base pair. In the first through the fourth panels, the black dots in Y-axis indicate read-depth signal from representative strains of each haplotype with their names and sequencing coverage labeled in the right margin; the red solid lines are the boundaries from GENSENG prediction; the green dashed lines are the boundaries reported by the Mouse Genomes Project (Clop et al., 2012); the grey lines are the median read-depth of the chromosome. The GC content and mappability of the region are plotted in the fifth and the sixth panels respectively. The last panel shows the locations of segmental duplication (purple) and repeat-mask repetitive DNAs (orange).The legend of Figure 2.11 is as follows. There are 6 panels from the top to the bottom. The X-axis of each panel indicates genomic position in base pair. In the first through the fourth panels, the black dots in Y-axis indicate read-depth signal from representative strains of each haplotype with their names and sequencing coverage labeled in the right margin; the red solid lines are the boundaries from GENSENG prediction; the green dashed lines are the boundaries reported by the Mouse Genomes Project (Clop et al., 2012); the grey lines are the median read-depth of the chromosome. The GC content and mappability of the region are plotted in the fifth and the sixth panels respectively. The last panel shows the locations of segmental duplication (purple) and repeat- mask repetitive DNAs (orange). Note that beginning at approximately 3684100 bp, there appears to be a deletion-like artifact in all strains with zero-valued read-depth. Close inspection suggested that it was caused by zero-valued mappability and was not called as deletion by GENSENG because of its ability to correct biases. Figure 2.10 shows a shared duplication from the mouse chromosome 17, also found from microarray studies (e.g. (Stranger et al., 2007)). For this region, AKR/J, C3H/HeJ, CBA/J, 129S1/SvImJ, A/J, DBA/2J, and LP/J belong to the same haplotype (Cahan et al., 2009) that contains this duplication, hence identity by descent. NOD/ShiLtJ and BALB/cJ belong to another haplotype (Cahan et al., 2009) that does not contain this duplication. Figure 2.11 shows a shared deletion from the mouse chromosome 17. Within the compatible interval with no historical recombination, chr17:36775200-36842845 (Cahan et al., 2009), AKR/J, C3H/HeJ, and CBA/J belong to the same haplotype (Cahan et al., 2009) that contains this deletion, hence identity by

Table 2.8: Summary GENSENG results from the mouse dataset

	Deletions				Duplications			
Inbred Mouse Strain	# GENSENG CNV calls (total Mbps)	# Yalcin CNV calls	# Yalcin CNVs overlapped (>1 bp) by GENSENG calls (percent)	# Yalcin CNVs overlapped (>50% overlap) by GENSENG calls (percent)	# GENSENG CNV calls (total Mbps)	# Yalcin CNV calls	# Yalcin CNVs overlapped (>1 bp) by GENSENG calls (percent)	# Yalcin CNVs overlapped (>50% overlap) by GENSENG calls (percent)
129S1SvlmJ	4387(24.8)	6262	1487(24%)	1474(24%)	428(4.7)	64	28(44%)	22(34%)
A/J	4931(24.9)	5674	1379(24%)	1366(24%)	64	23(36%)	20(31%)	
AKR/J	5961(34.7)	6007	2135(36%)	2123(35%)	1348(4.9)	78	36(46%)	33(42%)
BALB/cJ	4861(32.6)	5509	1450(26%)	1435(26%)	474(3.2)	73	26(36%)	22(30%)
C3H/HeJ	4827(29.4)	6035	1870(31%)	1849(31%)	1509(4.3)	84	26(31%)	22(26%)
CAST/EiJ	10081(55.3)	8995	679(8%)	637(7%)	351(8.8)	77	17(22%)	12(16%)
CBA/J	4579(28.2)	17544	876(5%)	797(5%)	575(3.4)	348	22(6%)	11(3%)
DBA/2J	5957(34.2)	6268	1270(20%)	1244(20%)	889(3.3)	63	19(30%)	14(22%)
LP/J	6551(34.2)	6513	1147(18%)	1128(17%)	736(5.0)	60	22(37%)	20(33%)
NOD/LtJ	11484(58.8)	6371	992(16%)	970(15%)	165(2.3)	60	19(32%)	11(18%)
NZO/HILtJ	6191(30.4)	5953	800(13%)	773(13%)	491(4.5)	45	12(27%)	11(24%)
PWK/PhJ	15780(73.4)	17901	5409(30%)	5390(30%)	348(4.7)	90	37(41%)	31(34%)
WSB/EiJ	6120(32.6)	5741	618(11%)	595(10%)	526(5.5)	56	16(29%)	13(23%)

descent. NOD/ShiLtJ, 129S1/SvImJ, A/J, DBA/2J, BALB/cJ, LP/J belong to another haplotype (Cahan et al., 2009) that does not contain this deletion. In addition, for the mouse strains, we compared the deletions and duplications predicted by GENSENG to those predicted by the Mouse Genomes Project (Yalcin et al., 2011; Simpson et al., 2009). We found that the overall concordance rates ranged from 3% to 4% (Table 2.8). A similar range of concordance was observed by the 1000GP by comparing CNV callsets generated by 19 algorithms. Furthermore, compared to the algorithms used by the Mouse Genomes Project (Yalcin et al., 2011; Simpson et al., 2009) we found that GENSENG had higher sensitivity for detecting duplications and comparable sensitivity for deletions (Table 2.8). The released SV calls (Yalcin et al., 2011) that are compared with have been classified into several categories based on specific paired-end mapping patterns. We extracted 2 categories, including deletions and copy number gains (GAINS and TANDEMDUP), to compare to the GENSENG predicted calls. This comparison focuses on the 19 autosomes, and GENSENG filter removes any predicted calls that have RDA values ranging between 0.5-2. We note that the improved sensitivity could be credited to GENSENGs bias-correction ability, which was absent in the approaches used by the Mouse Genomes Project (Yalcin et al., 2011; Simpson et al., 2009); this improved sensitivity warrants further experimental validation of the GENSENG-predicted duplications for the mouse strains.

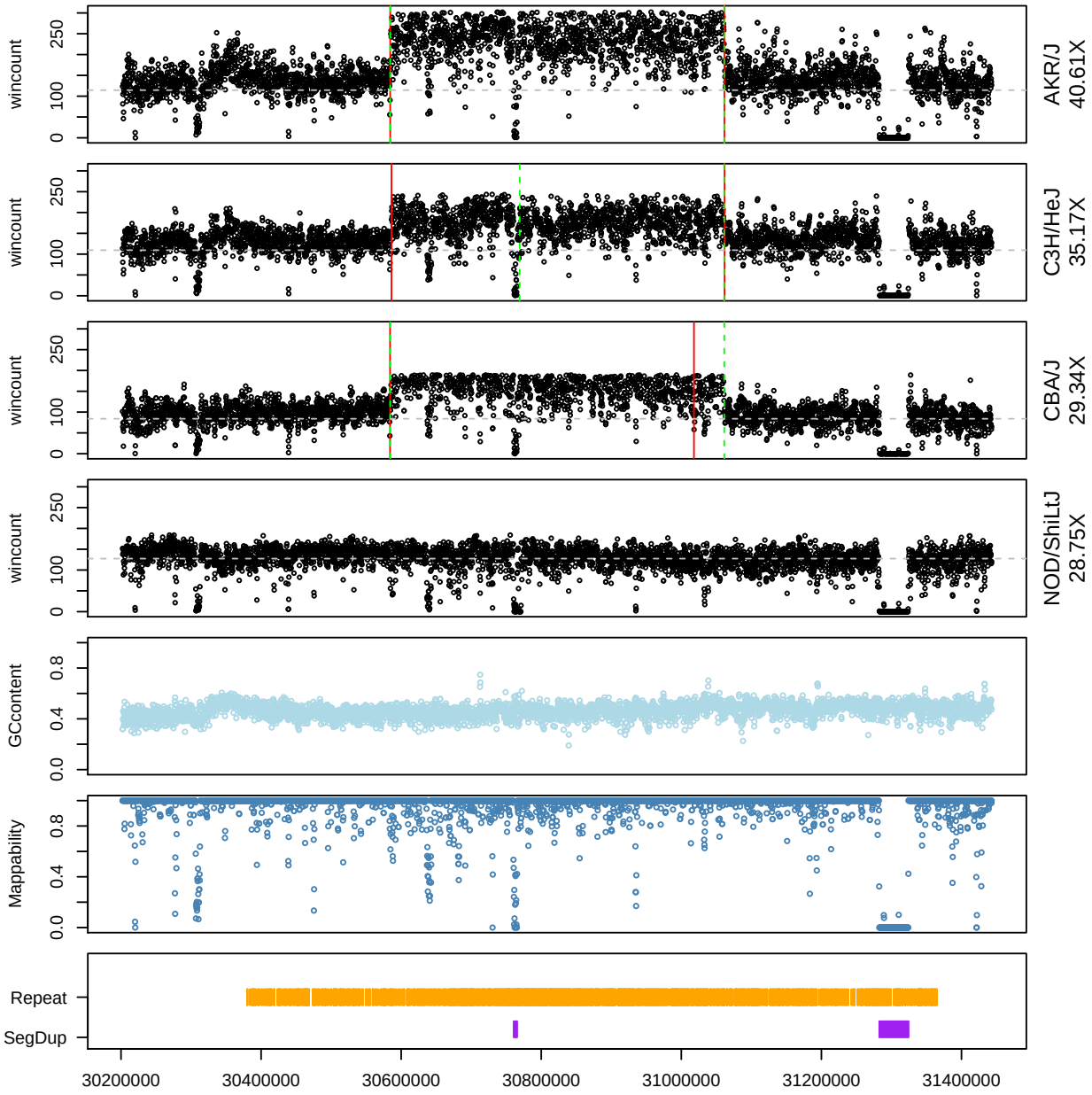


Figure 2.10: Example shared duplication identified from mouse HTS datasets.

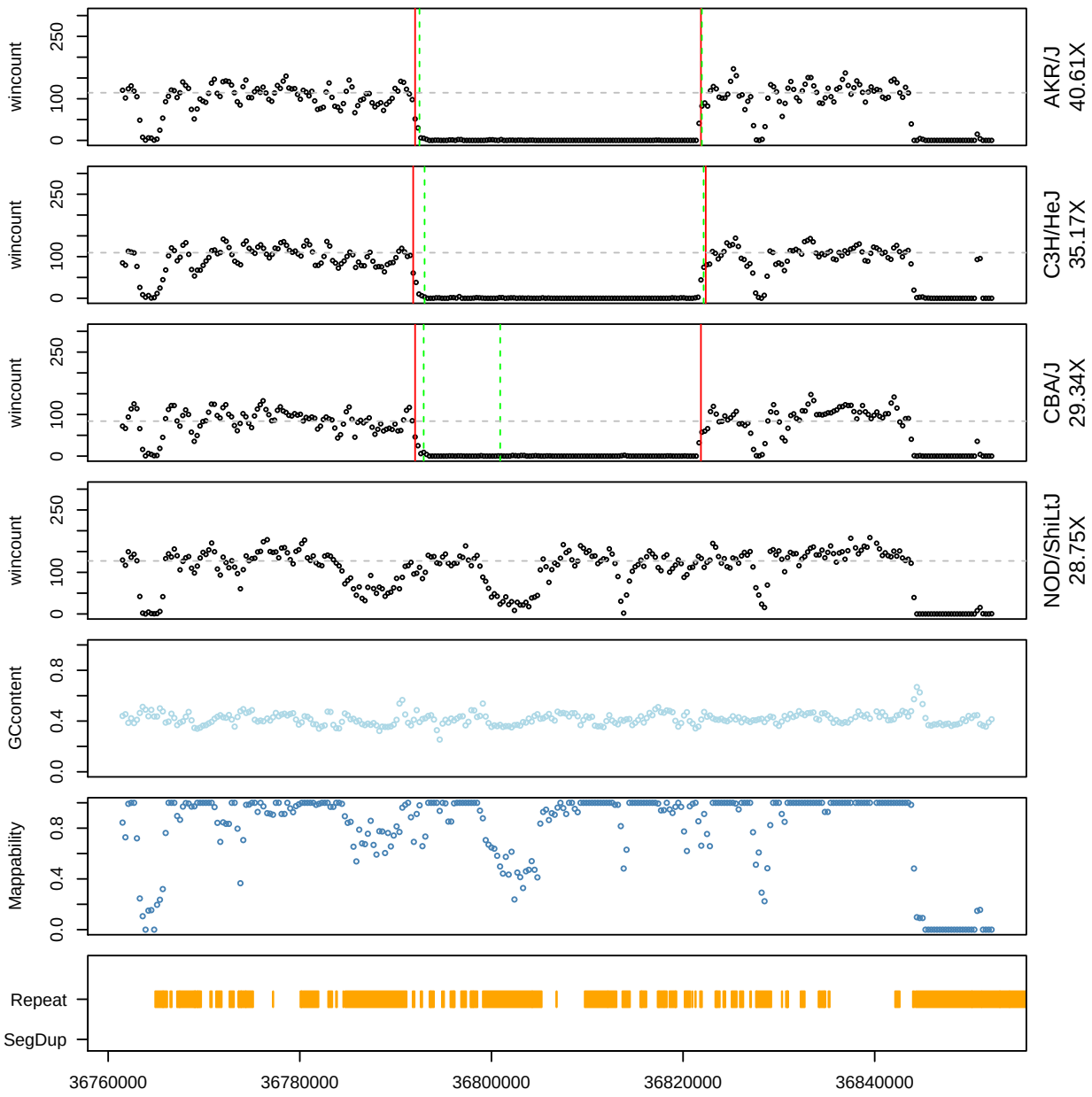


Figure 2.11: Example shared deletion identified from the mouse HTS dataset.

2.4 Discussion

We have developed a novel method, GENSENG (Szatkiewicz et al., 2013), for detecting copy-number gain and loss from HTS data. One unique feature and a key advantage of our method is the ability to simultaneously correct for multiple sources of bias and infer CNVs from read-depth. The concept of simultaneous bias correction and CNV inference can serve as a basis for combining read-depth with read-pair or split-read in a single analysis.

The GENSENG method can be applied to whole-genome sequencing data using either single-end or paired-end reads or a mixture of the two. It does not require matched control genomes, and it does not rely on evidence from multiple individuals. The smallest CNVs that can be detected by GENSENG are bounded by the window size applied in GENSENG (in Chapter 4, we will introduce an efficient version of GENSENG that facilitates the application to smaller window), and discrete copy numbers (0, 1, 2, 3, 4, 5, and 6+) are reported. Based on extensive benchmarking, GENSENG provides a better sensitivity-specificity profile than the previously best-performing read-depth-based algorithm, CNVnator (Mills et al., 2011; Abyzov et al., 2011) when applied to high-coverage HTS data. We have also demonstrated that our method works on both human and mouse samples with lower coverage (15x).

Our likelihood-based method can be readily extended to incorporate all the sequence reads and the mapping uncertainty. In addition, we can incorporate other types of information, such as haplotype, read-pair, split-reads, and allele-specific read depth that can infer allele-specific copy number. Allele-specific read depth can be informative for CNV calling. For example, in one window, if we observe approximately 50 reads from paternal allele and 100 reads from maternal allele, a reasonable guess is that the ratio of the number of maternal allele vs. paternal allele is 1:2, which will favor copy number 3 (1 paternal + 2 maternal), or copy number 6 (2 paternal + 4 maternal) etc. The allele-specific read depth can be incorporated as part of the emission probability, e.g. using a binomial distribution similar to the setup of the B-allele frequency following the genoCN method (Sun et al., 2009). In Chapter 3, we will introduce our method AS-GENSENG that integrated the allele specific information into the likelihood-based method.

CHAPTER 3: AS-GENSENG

3.1 Overview

As mentioned in Chapter 2, we have developed GENSENG (Szatkiewicz et al., 2013), a read-depth-based approach that accurately detects CNVs from WGS data by simultaneous read-depth segmentation and bias correction. Building upon the success of this integrative approach, the aim of this chapter was to develop an integrated method, named AS-GENSENG, that can (1) detect CNVs by jointly exploiting patterns in total- and allele-specific read count; (2) estimate ASCN; and (3) be applicable to both WGS and WES data. For bias correction, we inherited the one-step approach used by GENSENG (Szatkiewicz et al., 2013) and leveraged allele-specific information for normalizing WES data.. Here, we evaluate AS-GENSENG using simulation and WGS or WES data from the 1000 Genomes Project (1000GP) (Mills et al., 2011; Abecasis et al., 2012) and compare our method to a number of state-of-the-art CNV detection algorithms in the literature (Heinzen et al., 2012; Fromer et al., 2012; Plagnol et al., 2012; Krumm et al., 2012; Abyzov et al., 2011). Furthermore, we validate a subset of CNV calls with an independent and highly accurate technology (NanoString nCounter) (Geiss et al., 2008; Sailani et al., 2013; Iskow et al., 2012; Ruderfer et al., 2013; Brahmachary et al., 2014). In summary, we conclude that AS-GENSENG not only predicts accurate ASCN calls, but also improves the accuracy of total copy number calls. For WGS data, AS-GENSENG has better overall performance in detecting CNVs than several state-of-the-arts methods for WGS data. For WES data, AS-GENSENG has better sensitivity and comparable specificity for detecting common CNVs from WES data. Our novel, user-friendly and computationally efficient method is freely available at `https://sourceforge.net/projects/asgenseng/`.

3.2 Materials and Methods

3.2.1 Method summary

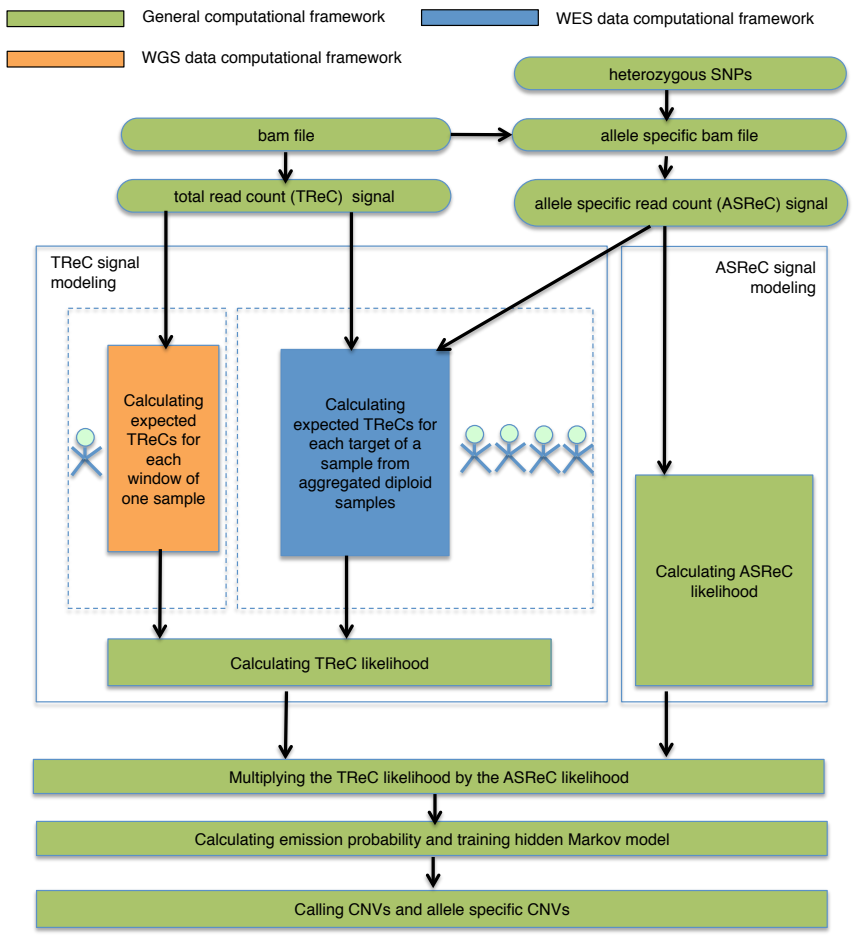


Figure 3.1: AS-GENSENG overview

High-throughput sequencing (HTS) captures multiple sources of information in one experiment, including read-depth, read-pair, split-read, and allele-specific read count information. Inspired by the successful integration of probe intensity and SNP genotype data in array-based CNV calling, here we develop an analogous method using HTS data. By incorporating

Table 3.1: Allelic configuration correspondences for each HMM state

	Allelic Configurations (expected allelic imbalance ratio from allele A) correspondence for each HMM state[1]						
Interpretation	0 Homozygous deletion	1 Heterozygous deletion	2 Copy number neutral	3	4	5 Duplications	6+
Allelic configurations (expected allelic imbalance ratio)	-	A (0.99)[2]	AB (0.5)	ABB (0.33)	ABBB (0.25)	ABBBB (0.2)	ABBBBB (0.17)
		B (0.01)[2]		AAB (0.67)	AABB (0.5)	AABBB (0.4)	AABBBB (0.33)
					AAAB (0.75)	AAABB (0.6)	AAABBB (0.5)
						AAAAB (0.8)	AAAABB (0.67)
							AAAAAB (0.83)

allele-specific read count information with read-depth data (total number of reads in each genomic region), our method AS-GENSENG can accurately detect CNVs and allele-specific CNVs (ASCNV) from both WGS and WES data. Figure 3.1 provides an overview of our method. First, AS-GENSENG jointly exploits patterns in Total Read Count (TReC) and Allele-Specific Read Count (ASReC) signals. TReC is analogous to the total intensity measured from SNP arrays. To detect CNVs, we look for higher or lower TReC than the expected TReC while simultaneously accounting for various sources of bias. As with GENSENG (Szatkiewicz et al., 2013), multiple known biases (e.g. GC content and mappability) are corrected explicitly using a covariate method, while unknown biases are accommodated by the over-dispersion parameter of the negative binomial distribution and by an additional noise component via a mixture model. ASReC is analogous to the allelic intensity measured from SNP arrays, with expected patterns of allelic imbalance for each copy number state (Table 3.1). For example, Figure 3.2 shows a region of copy number four (enclosed by vertical lines) flanked by regions with copy number two. After taking into account sources of bias (Figure 3.2c), the TReC in the enclosed region is approximately two times higher than that of the flanking region (Figure 3.2a), which supports a duplication with copy number four. While the proportion of ASReC from allele A in the two copy region are $\sim$0.5, this proportion is $\sim$0.25 in the enclosed region, supporting a call of copy number four with an allelic configuration of ABBB (Figure 3.2b). Figure 3.3 shows an example duplications, and Figures 3.4,3.5 show examples of copy number one, where both TReC and ASReC deviate over the CNVs in comparison to the flanking copy number two regions. These observations suggest that jointly exploiting patterns in both TReC and ASReC should improve the ability to detect both deletions and duplications.

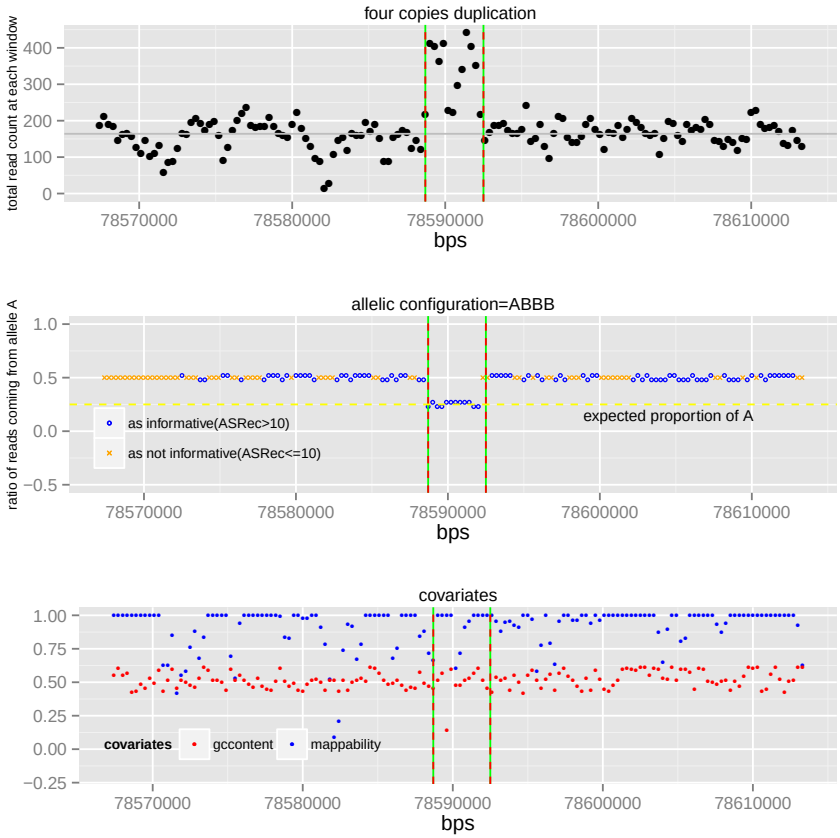


Figure 3.2: Jointly analysis by AS-GENSENG

Furthermore, ASReC is useful for detecting common CNVs (i.e. CNVs >5% frequency) from WES data because it accurately identifies the reference copy number two group without prior assumption of CNV frequency. Figure 3.6 shows an example of this phenomenon for a common CNV using 1000GP WES data (Mills et al., 2011; Abecasis et al., 2012) In this example, >40% of the samples have a deletion that is either homozygous (copy number 0) or heterozygous (copy number one). Typically, read-depth-based approaches examine observed read counts (observed TReC) versus the read count expected for copy number two (expected TReC).

---

[1] In this study, we assume there is no complex CNV (i.e., that no two CNVs occur at the same genomic location). Thus, for example, for state 3, we define 2 allelic configurations, ABB and AAB, as being possible only with duplication. We do not define AAA or BBB, because they require both deletion and duplication to occur at the same location.

[2] A very small amount of reads will be aligned to a genomic region even when that region has been deleted. Thus we assume that if allele A is kept and allele B has been deleted, not all reads would show allele A, and vice versa. Thus the expected allelic-imbalance ratio from allele A of allelic configuration A is 0.99, not 1.0, and the expected allelic imbalance ratio from allele A of allelic configuration B is 0.01.

Higher-than-expected TReC indicates a duplication, whereas lower-than-expected TReC indicates a deletion. At least two approaches were developed to identify the reference group of copy number two and estimate the expected TReC using the reference. The first approach (Li et al., 2012) uses the median or trimmed mean of all samples to estimate the expected TReC of the copy-number-two reference, assuming that most samples have copy number two at a target (Nord et al., 2011; Plagnol et al., 2012; Li et al., 2012). However, as the median of all samples is far from the median of the copy-number-two reference group in the presence of common CNVs (Figure 3.6 (a)), the median approach would lead to incorrect inference of the underlying copy numbers. The second approach (Plagnol et al., 2012) constructs an optimized copy number two reference set by ranking the correlations of TReC between the reference and the test exome and assuming that the CNV call is not present in the reference exomes. However, empirical results suggested that this approach had limited power for detecting common CNVs, presumably because the no-CNV assumption does not always hold in the selected reference exomes (Plagnol et al., 2012). By contrast, as shown in Figure 3.6 (b), we used ASReC and allelic imbalance ratio differences to first properly identify a reference group with copy number two. Then, given the accurate estimation of expected TReC for copy number two, the clusters of samples with copy numbers one or zero can be easily identified by the joint use of TReC and ASReC for this target.

3.2.2 Data preparation

Our AS-GENSENG method requires as input aligned sequence data (and a subset of extracted sequence alignments that carry allele specific information). AS-GENSENG counts the read-depth from these alignments and analyses the read-depth for CNV detection. In order to calibrate the CNV detection performance of AS-GENSENG, we used published data as a comparator, as detailed below.

Alignment files. We used whole-genome sequencing (WGS) data from two HapMap individuals and whole-exome sequencing (WES) data from 324 HapMap individuals sequenced as a part of the 1000 Genomes Project (1000GP) (Mills et al., 2011; Abecasis et al., 2012). The WGS samples included two HapMap samples of European ancestry (NA12891, NA12892), sequenced to

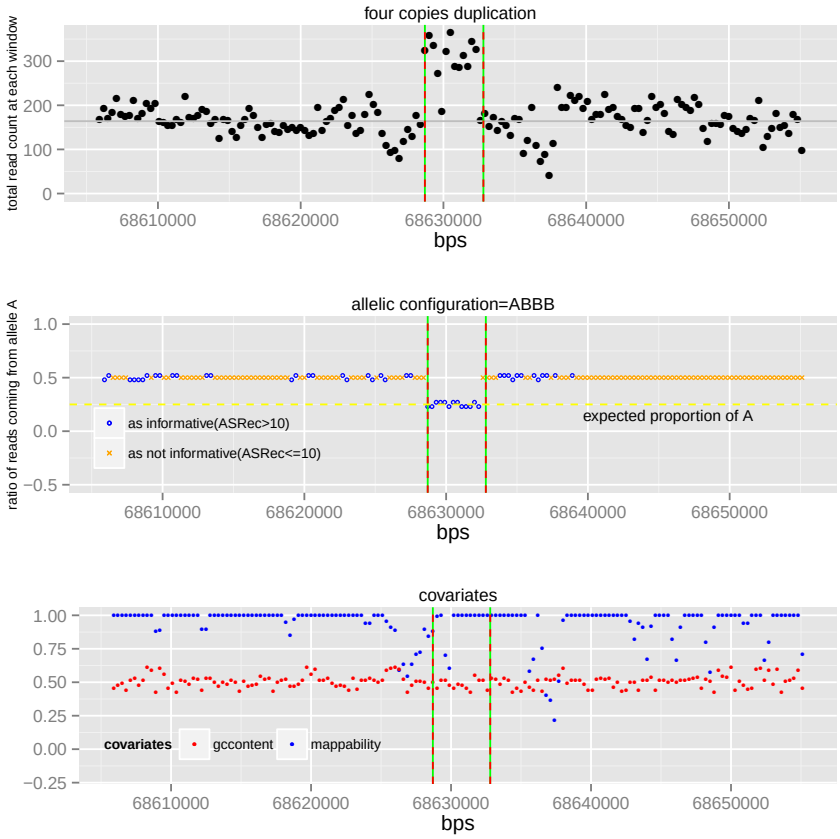


Figure 3.3: AS duplication example

an average depth $<$ 40X using the Illumina Genome Analyser (I and II) platform. Sequencing reads were a mixture of single-end and paired-end, and of variable read-length (36bp or 51bp). Reads were aligned using BWA (Li and Durbin, 2009) (v0.5.5); the reference was NCBI37 (see `ftp://ftp.ncbi.nlm.nih.gov/1000genomes/ftp/README.alignment_data` for details). The WGS data were obtained as .bam alignment files (`ftp://ftp.ncbi.nlm.nih.gov/1000genomes/ftp/pilot_data/data/`). The WES samples include 324 individuals from four different populations (European, Asian, American, and African) sequenced to an average depth of $\sim$100X using different capture platforms: SeqCap EZ Human Library (v1.0 and v2.0) from Nimblegen and SureSelect All Exon V2 Target Enrichment kit from Agilent. A consensus-capture target-region list is first defined by intersecting all the target design files (.bed files) with the NCBI CCDS database, and then adding 50 bps at either side of each consensus target. The list contains 193,637 exome capture targets and $\sim$47Mbps

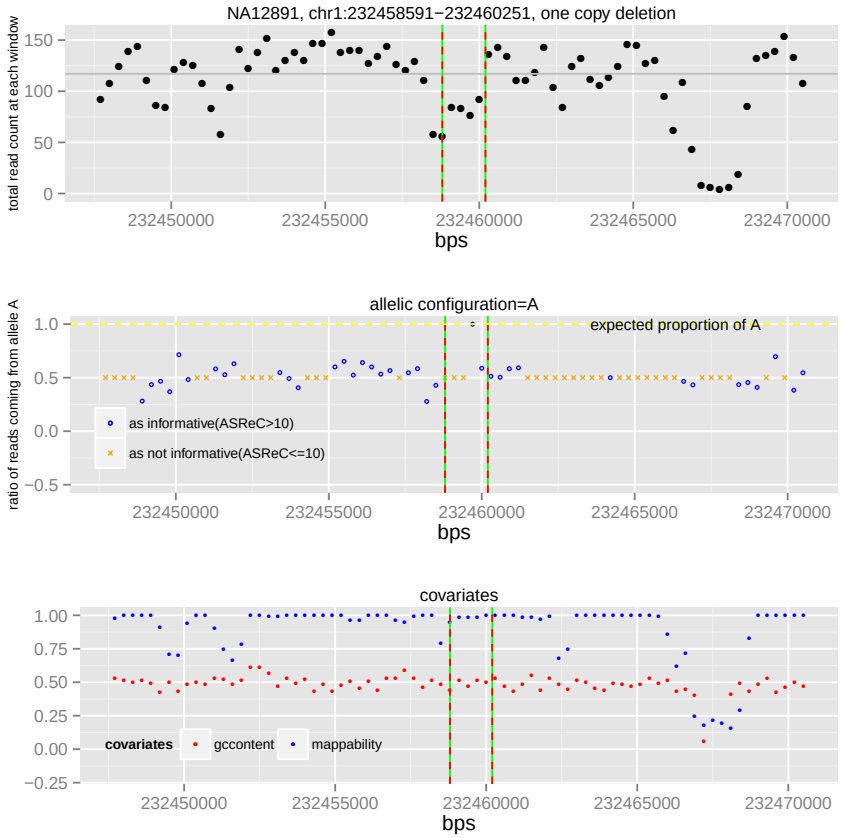


Figure 3.4: AS deletion in NA12891 recovered in AS-GENSENG

are captured (listed at `ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/exome_pull_down_targets/20130108.exome.targets.bed`). Sequence reads were all paired-end and read-length was uniform within a sample, but varied among samples (76bp, 90bp, and 100bp). Reads were aligned using BWA (Li and Durbin, 2009)(v0.5.9) to NCBI37 (see `ftp://ftp.ncbi.nlm.nih.gov/1000genomes/ftp/README.alignment_data` for a detailed description). The WES data were obtained as .bam alignment files (`ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/data/`).

Heterozygous SNPs. AS-GENSENG requires dense, phased genotypes to make proper allele specific CNV calls (alleles A and B must be consistent across different markers). SNP genotypes were first obtained from microarray or sequence-based SNP-calling algorithms, such as samtools (Li et al., 2009) or GATK (McKenna et al., 2010; DePristo et al., 2011). We then carried

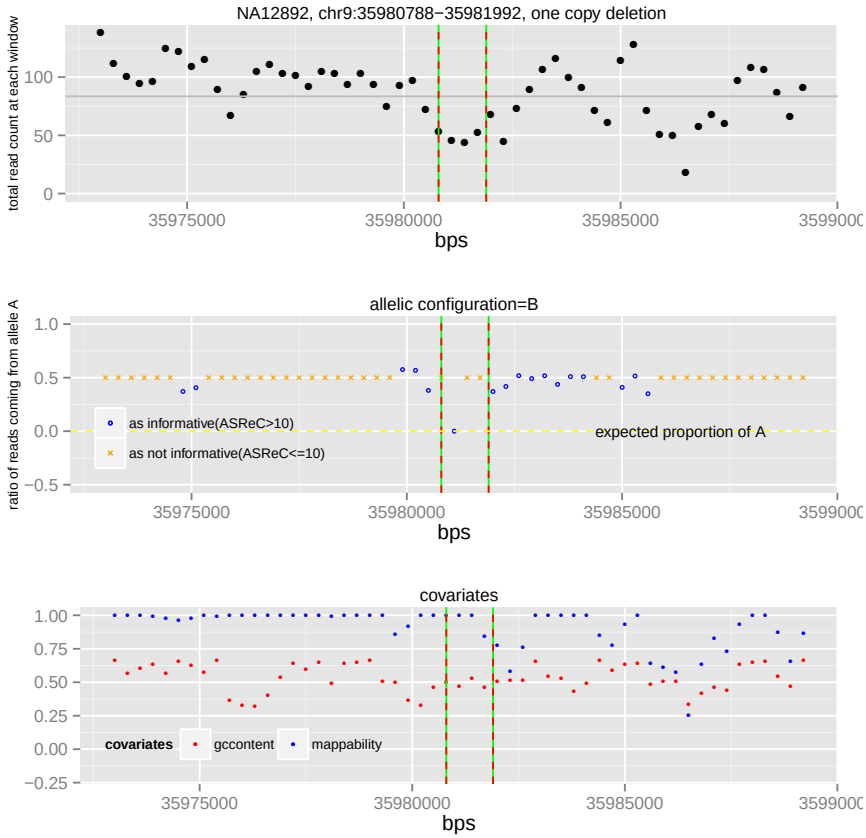
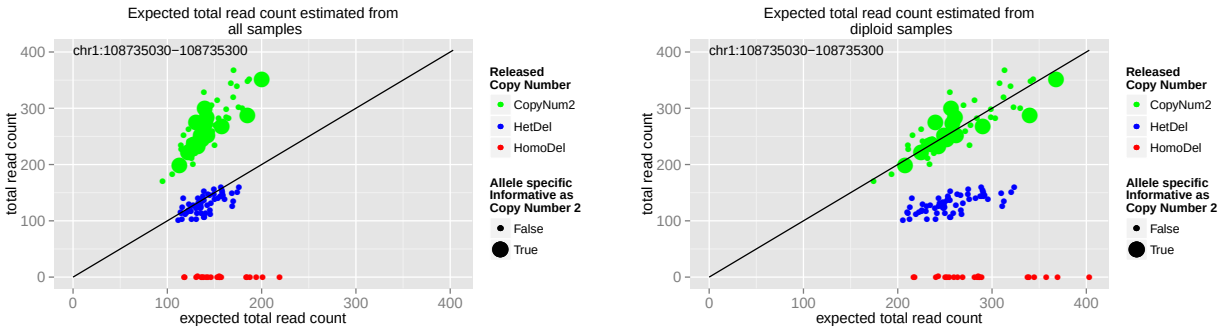


Figure 3.5: AS deletion in NA12892 recovered in AS-GENSENG

out imputation to obtain a phased and dense list of input SNPs. For the two studied WGS samples, we used MaCH-Admix (Liu et al., 2013) to impute from the relatively sparse HapMap3 r2 SNPs list (obtained from microarray `http://hapmap.ncbi.nlm.nih.gov/downloads/phasing/2009-02_phaseIII/HapMap3_r2/CEU/TRIOS/`) to generate a dense-phased SNP list. The imputation for the reference SNP panel was 1000 GP-released PhaseI V3 haplotypes on 1092 samples (phased, downloaded from `ftp://share.sph.umich.edu/1000genomes/fullProject/2012.03.14/phase1_release_v3.20101123.snps_indels_svs.genotypes.refpanel.ALL.vcf.gz.tgz`). The SNP coordinates reported in HapMap3 r2 were translated from NCBI36 to NCBI37 using liftOver; the reference panel was NCBI37. After imputation, we extracted 198k and 200k phased heterozygous SNPs from samples NA12891 and NA12892, respectively. For the 324 WES samples, we used an in-house



(a) Expected RC using all samples

(b) Expected RC using CN two samples

Figure 3.6: Example: Common Deletion in WES data

python script to extract phased heterozygous SNPs from the reference panel (on average, 2.2 million heterozygous SNPs per sample). The coordinates were NCBI37.

Allele-specific alignment files. Allele-specific alignments (AS alignments) are aligned reads that could be confidently assigned to one particular SNP allele. We used the function extractAsReads in R/asSeq (`http://www.bios.unc.edu/~weisun/software/asSeq.htm`) to extract AS alignments passing QC from alignment files. For each aligned read, the asSeq package searched the heterozygous SNP list by coordinates and counted the number of SNPs from alleles A or B that the read carried. If the count of allele A was >0, asSeq output the read to the A allele file, and if the count of allele B is >0, asSeq output it to the B allele file. If both counts were > 0, the read was treated as an error, because each read is expected to carry only one allele. Such reads were discarded, but they were rare ($<0.1\%$) in our experiment.

Total and allele-specific read counts. In each genomic region, we counted the total number of reads aligned (TReC) and the number of AS alignments (AsReC). Although each region has only one TReC value, the ASReC in each region consists of two values: $o^{(A)}$ (i.e., the number of A allele reads) and $o^{(as)} = o^{(A)} + o^{(B)}$ (i.e., the number of A allele reads plus B allele reads). Note that $o^{(as)}$ is smaller than TReC because many reads do not overlap a heterozygous SNP. We first applied a quality control (QC) procedure using SAMtools (Li et al., 2009) to extract confidently aligned reads from both the TReC and AS alignment files (see Section 2.2.2 for QC details). All subsequent

analyses were based on reads that passed QC. Because a sequence read represents one or two ends of a DNA fragment randomly sampled from the genome, we treated a sequenced DNA fragment as the counting unit and ensured that each fragment was counted only once. For WGS data, we used 500bp-sliding windows with a step size of 100bp and this choice was first determined via simulation and then verified empirically. In our simulation study, we first simulated sequencing reads from a hypothetical CNV-containing genome (see Section 3.2.5 for detail), and then computed four sets of read-depth data for various sizes of sliding windows: 200bps, 300bps, 400bps and 500bps (each with a step size of 100bp). The best detection performance was achieved for 500bp-windows (93% sensitivity and 4% FDR). When 200bp-windows were used, we observed slightly improved sensitivity (94%) but much higher FDR (9%). This result was verified by empirical experiments, where we applied AS-GENSENG to 1000GP WGS data using both 500bp- and 200bp-sliding windows. Results for 500bp-windows are reported in Table 3.7 of the main paper. When 200bp-windows were used, we observed higher sensitivity ($\sim$3% higher for each HapMap sample analyzed in this study), but much reduced specificity ($>$20% more predicted CNV calls). Therefore, both simulation and real-data analysis confirmed the choice of 500bp-sliding windows used in this study, presumably because it resulted in better signal-to-noise ratios in comparison to smaller size windows. We calculated TReC as the number of DNA fragments in each window (see Section 2.2.2 for counting details). For WES data, we calculated TReC as the number of DNA fragments in each captured target using the following procedure.

1. Characterize the fragment as the left-most aligned position and the right-most aligned position of the two paired ends.

2. Check the quality of the characterized fragment. Remove if it looks abnormal, e.g. the length of its spanned region is much larger than the library insert size.

3. Intersect the spanned region with the targets list. Assign 1 to the overlapped target if it only intersects with one target. Assign $1/N$ to overlapped targets if it intersects with N targets.

ASReC was counted similarly, except the input was extracted allele-specific reads. We counted $o(A)$ and $o(B)$ for each window using the AS alignments for alleles A and B. We then summed the two values at each window to obtain $o^{(as)}$.

Covariate files. Local GC content and mappability create variability in TReC (Szatkiewicz et al., 2013). GC content was computed as the proportion of GC bases in each genomic region (200bps windows in WGS data or a capturing target in WES data) of the reference genome. Mappability was computed as the proportion of bases in each genomic region of the reference that could be uniquely aligned to using the minimal read-length of the data (covariate details given in Supplemental Methods).

High-confidence CNV data. To evaluate the sensitivity of our method, we used high-confidence CNVs previously generated from the same samples that we studied. For the two high-coverage WGS samples (NA12891, NA12892), we used the high-confidence deletions established by 1000GP (Mills et al., 2011; Abecasis et al., 2012; Handsaker et al., 2011), downloaded from `ftp://ftp.broadinstitute.org/pub/svtoolkit/misc/1kg/NGPaper/`. This dataset had been validated as having the lowest false discovery rate (FDR $< 4\%$) in the 1000GP (Mills et al., 2011; Abecasis et al., 2012) and therefore was considered to be the best available high-confidence CNV data for these samples. This dataset included 2200 deletions for NA12891 and 2055 deletions for NA12892. CNV coordinates reported in this dataset were translated from NCBI36 to NCBI37 using liftOver. For the 324 HapMap samples for which we obtained WES data, WGS data was also available from the 1000GP (Mills et al., 2011; Abecasis et al., 2012). Using the WGS data, high-confidence genome-wide deletions have been established and released at `ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521/`. These genome-wide deletions were validated by multiple independent technologies as having high specificity (FDR $< 10\%$) by 1000GP (Mills et al., 2011; Abecasis et al., 2012). Similar to (Tan et al., 2014),we intersected this genome-wide dataset with the exon target list using a 1bp reciprocal overlapping condition to obtain high-confidence deletions in the target regions (i.e.

exonic-deletions) as a comparator set for sensitivity evaluation. In total, there were 9,192 exonic-deletions for the 324 WES samples. As WGS is a more powerful technology in identifying CNVs than WES, these high-confidence exonic-deletions provide both validity and accuracy in evaluating exonic CNVs identified by WES (Tan et al., 2014).

Similar to the definition in Section 2.2.2, the input data is a sequence of two-tuples for each studied genomic region (window, or target) represent by

$\{O, X\} = \{o_1.....o_T, x_1.....x_T\}$, where T is the total number of genomic regions studied of a chromosome. $o_t = (o_t^{all}, o_t^{(A)}, o_t^{as})$ denotes the TReC, ASReC from allele A, and ASReC from both alleles of the t^{th} genomic region. $x_t = (g_t, l_t)$ denotes the covariates of the t^{th} genomic region, where g_t represents the GC content, and l_t denotes the mappability score.

3.2.3 Hidden Markov model

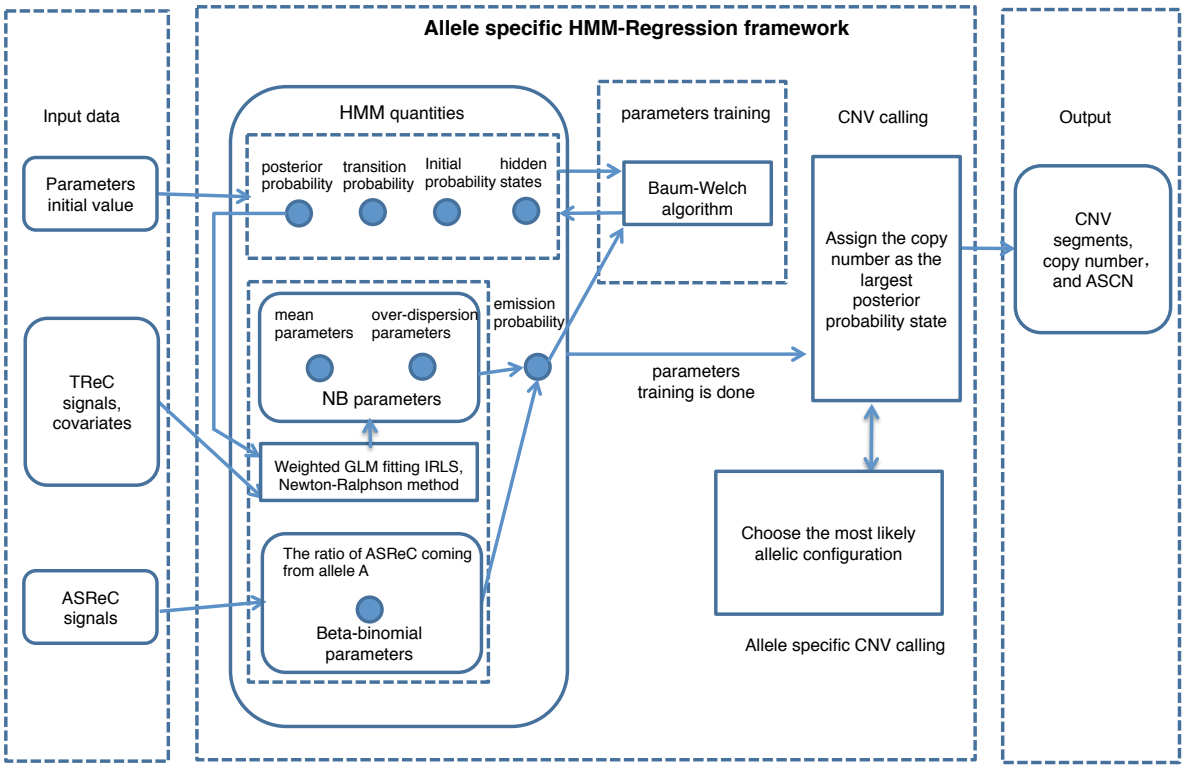


Figure 3.7: HMM flowchart to infer ASCN

We developed a hidden Markov model (HMM) classifying each genomic region (a window for WGS data or a WES target) to a copy-number state based on maximum a posteriori probability. In comparison to other segmentation approaches such as circular binary segmentation (Olshen et al., 2004), the use of HMM allows the joint analysis of multiple sources of information (TReC, ASReC, covariate values) as well as the modelling integer copy numbers. Our HMM consists of multiple components as shown in Figure 3.7. The main procedure includes: 1) input data consisting of TReC, ASReC, and covariates at each window or target; 2) an emission probability jointly modelling both TReC and ASReC, and accounts for confounding factors such as GC content or mappability; 3) an EM algorithm (Weiss et al., 1970) optimizing model parameters to find the

optimal underlying copy number for each region, with the state changes indicating the predicted CNV breakpoints. 4) in the final step, the model calls ASCNV. Below we first present a brief description of each step. The main difference with the model in Section 2.2.4 is the integration of allele specific information. In the following paragraphs, we will highlight the difference with GENSENG. Finally, to highlight an important contribution of AS-GENSENG, we introduce a model for inferring ASCN in WES emission probability data in the following section.

HMM state. Similar to Section 2.2.4, the total number of hidden states is an input parameter and can be specified by users. For the datasets used in this study, we set seven hidden states that respectively represent copy numbers of 0, 1, 2, 3, 4, 5, and 6 or more. In this work, the duplications with 6 or more copies were collapsed into one state because they were difficult to differentiate. To model ASCN, we defined several possible allelic configurations for each state (Table 3.1). For example, we defined AAB and ABB as the two possible allelic configurations for the one copy duplication (i.e., state 3).

Transition probability. We model the state transitions using a first-order time-homogeneous Markov process (i.e., the state in one genomic region is affected only by the immediately previous region). Under this setting, the transition probability describes the probability of having a copy number change between two adjacent genomic regions. The transition probability is characterized as a square matrix, of which the dimension is the number of states and the (i, j) element is the probability of transition from state i to state j. We set the transition probability matrix according to our intuition that the copy number state is unlikely to change for nearby genomic regions but is likely to change for genomic regions that are far apart. Thus the self-transition probability (i.e., the diagonal values of the matrix) is much larger than the transition probability of transiting to other states. We assumed that most windows would have copy number two. Thus, the self-transition probability for state two would be higher than that of other states. In addition, the probability of transiting to state two would also be larger than the probability of transiting to other states. To handle the problem of varying distance between targets in WES data, we further modified each element in the transition matrix as suggested in Fromer et al (Fromer et al., 2012). The new element

for the (i, j) element $a^{'}(i, j)$ would be a mixture of two original elements at (i, j) and $(2, j)$, as $a^{'}(i, j) = e^{-d/D}a(i, j) + (1 - e^{-d/D})a(2, j)$, where d is the distance between two targets and D is the average distance between all targets.

Emission probability. Emission probability specifies the likelihood of observing the TReC inputs and the ASReC inputs given the underling copy number and other information at the region (i.e., the covariates), which could be expressed as $Pr(o_t = (o_t^{all}, o_t^{(A)}, o_t^{as})|q_t = j, x_t)$, where $o_t^{all}, o_t^{(A)}, o_t^{as})$ are the TReC and ASReC, q_t is the underline hidden copy number, and x_t is the covariates at this genomic region. Given the underlying state, TReC and ASReC are independent and thus the likelihood can be factorized.

$$
\begin{aligned}
&Pr(o_t = (o_t^{all}, o_t^{(A)}, o_t^{as})|q_t = j, x_t) \\
&\quad = Pr(o_t^{all}|q_t = j, x_t)Pr(o_t^{(A)}|o_t^{as}, o_t^{all}, q_t = j, x_t)Pr(o_t^{as}|q_t = j, o_t^{all}, x_t).
\end{aligned}
$$

The first factor $Pr(o_t^{all}|q_t = j, x_t)$ is the probability of observing the TReC given the covariates and underling states, and the second factor $Pr(o_t^{(A)}|o_t^{as}, o_t^{all}, q_t = j, x_t)$ is the probability of observing the ASReC from allele A ($o^{(A)}$), given the overall ASReC and underlying states. For the third factor $Pr(o_t^{as}|q_t = j, o_t^{all}, x_t)$, given TReC o_t^{all}, the ASReC o_t^{as} is conditional independent of the state j and the covariates x_t. Thus $Pr(o_t^{as}|q_t = j, o_t^{all}, x_t) = Pr(o_t^{as}|o_t^{all})$. $Pr(o_t^{as}|o_t^{all})$ depends on the number of heterozygous SNPs in the t^{th} genomic region. In this study we modeled it as a constant.

For the first factor, the basic idea is that we would first obtain the baseline TReC of copy number two and then compare the observed TReC with the baseline to infer the copy number. The specific modelling approaches are different between WGS data and WES data, however. For WGS data, following Section 2.2.4, the TReC is modelled by a negative binomial distribution (NB). Known sources of bias such as GC content and mappability are included as the covariates of NB regression. Unknown sources of bias are accommodated by the NB overdispersion parameter and

the uniform distribution (Szatkiewicz et al., 2013). The method aggregates TReCs from all windows of one sample to estimate the expected TReC for copy number two and other copy numbers, with the assumption that the TReC would be proportional to the underlying copy number. The overdispersion parameter is estimated from the data using Newton-Raphson method. Below we describe a novel model for WES data incorporated into AS-GENSENG.

The second factor, which is the probability of observing the ASReC from allele A ($o^{(A)}$) given the overall ASReC ($o^{(as)}$), describes the allelic imbalance, analogous to BAF in SNP array data. In both WGS and WES, it is modelled by a beta binomial distribution (BetaB), which is an extension of a binomial distribution to allow for possible overdispersion (Sun, 2012). Specifically, let $o^{(A)}$ follow a binomial distribution with the number of trials $o^{(as)}$ and the probability of success p_s. If p_s follows a beta distribution with parameters α and β, the resulting distribution of $o^{(A)}$ is a beta-binomial distribution. This method adapts a commonly used strategy to parameterize a beta-binomial distribution by $\pi = \alpha/(\alpha+\beta)$ and $\theta = 1/(\alpha+\beta)$. Thus the likelihood of a beta-binomial distribution becomes:

$$\ell(o^{(A)}, o^{(as)}, \pi, \theta) = \binom{o^{(as)}}{o^{(A)}} \frac{\prod_{k=0}^{o^{(A)}-1}(\pi + k\theta)\prod_{k=0}^{o^{(as)}-o^{(A)}-1}(1-\pi+k\theta)}{\prod_{k=1}^{o^{(as)}-1}(1+k\theta)},$$

where π is the expected proportion of AS reads from allele A (e.g. $\pi = 0.33$ for allelic configuration ABB). θ is a dispersion parameter. If there is no over dispersion, then $\theta = 0$ and $o^{(A)}$ follows a binomial distribution. In this work, we empirically set $\theta = 0.1$. An underlying copy number has several possible allelic configurations (Table 3.1). We thus formulate the likelihood as the likelihood of a mixture distribution across all possible allelic configurations (e.g., for copy number 3, the likelihood would be $(\ell(o^{(A)}, o^{(as)}, 0.33, 0.1) + \ell(o^{(A)}, o^{(as)}, 0.67, 0.1))/2$).

Finally, due to the large observed noise in the data, we add a uniform distribution into the emission probability, so that the emission probability becomes:

$e(t,j) = c/R + (1-c)Pr(o_t^{all}|q_t = j, x_t)Pr(o_t^{(A)}|o_t^{as}, q_t = j)$, where $e(t,j)$ is the emission probability of the genomic region t given the underlying copy number $q_t = j (0 \leq j \leq 6)$;

c is the proportion of random uniform component which is constant for all states; R is the maximum read count; $1/R$ is the uniform density; $(o_t^{all}, o_t^{as}, o_t^{(A)})$ is the input observations tuple representing the TReC, total ASReC, and ASReC from allele A of the genomic region t, respectively; and x_t is the input covariates for the genomic region t. For datasets used in this study, c was set at 0.01 and was determined empirically by initializing the model with varying values of c and identifying the maximizer of the data likelihood.

HMM training and inference of total copy number. HMM training provides the maximum-likelihood estimate of the HMM parameters. To improve computational efficiency, transition-probability parameters were specified using prior knowledge and user preference (Szatkiewicz et al., 2013), and emission probability parameters were estimated using the Baum-Welch algorithm (Weiss et al., 1970). Using the estimated parameters, we compute the posterior probability of each genomic region belonging to a particular state and assign the most likely state for each region. The confidence score is computed as the sum of the posterior probabilities in regions spanned by a CNV.

Inference of ASCN. We assign the most likely ASCN given the most likely copy number call. For example, if the most likely copy number for a CNV is 3, AS-GENSENG chooses between ABB and AAB. It first selects windows with ASReC larger than a threshold (10 in this study) in the region and computes the average ASReC of the selected windows. If no window were selected, we would not infer ASCN because the ASReC is not informative. Otherwise, we would compute the likelihood of AAB and ABB using BetaB distribution introduced above and choose the one with the largest likelihood as the inferred ASCN.

3.2.4 Inference of ASCN in WES data

The input data (TReC and ASReC) are the same as in the WGS setting. In the WES setting, however, exome capture and sequencing results in non-uniform TReC signal between captured targets. Therefore, we expect TReC to differ between targets even if the underlying copy number is the same; so we cant apply the WGS approach to WES data in order to calculate expected TReC. Instead, we analyze in a target-by-target manner to estimate the expected TReC. We first normalize

sample TReC by computing the ratio between TReC at one target and the sum of TReC of all targets. If a given target has copy number two in most samples, we could use the median/or a trimmed mean (Li et al., 2012) of ratios from all samples to obtain the expected ratio of copy number two. However, when a common CNV exists, the median ratio is no longer a good estimator of the expected ratio of copy number two. We observe that those ratios in fact form a few clusters, with each cluster corresponding to a particular copy number (Figure 3.6). Therefore, we needed to develop a method of finding the proper reference group of samples that are likely to have copy number two, and use the median ratios of the reference group to estimate the expected ratio of copy number two. Here, we use ASReC information to identify the copy number two reference group. For each target, we first use the R package Mixtools (`http://cran.r-project.org/web/packages/mixtools/mixtools.pdf`) to find the number of sample clusters, and then generate clusters of samples based on the ratio value of each sample. We use ASReC to compute the probability of having copy number two for each cluster as follows: for each sample in one cluster, we compute its probability of being copy number two by dividing the BetaB likelihood of being copy number two with the sum of Beta likelihoods of being each copy number. The probability of a cluster being copy number two is the average probability of being copy number two for all AS informative samples in the cluster. We choose the cluster with the largest posterior probability of being copy number two, and use the median ratio value of this cluster to compute the expected TReC for the reference group of copy number two.

It must be pointed out that the procedure above is only used to compute the expected TReC for the reference group of copy number two and we do not assign copy number states simply based on this step, as there remains other information, such as TReC and information of neighbour targets, that is not yet utilized. We incorporate the expected TReC into the HMM inference framework (as above) to make the full use of all the available information.

3.2.5 Performance Evaluation: WGS data

We used both simulation and empirical data to assess the performance of AS-GENSENG in detecting CNVs and ASCNVs in WGS data. To compare performance, we applied both

Table 3.2: Methodologies comparisons among WGS methods

	AS-GENSENG	GENSENG	CNVnator	ERDS
Allele-specific CNV	Y	N	N	N
Analyzing exome sequencing data	Y	N	N	N
Using read-pair information	N	N	N	Y for <10kbps deletions
Bias correction	1-step approach Correct for GC content, mappability, and additional noise in the data.	1-step approach. Correct for GC content, mappability, and additional noise in the data.	2-step approach. Correct only for GC content.	2-step approach.Correct only for GC content.
Use of allele-specific information	Use of beta-binomial distribution to model the allelic imbalance, and combine it into emission probability of HMM.	N	N	Modeling the total number of heterozygous SNPs in each window, and combining it into the emission probability of HMM.
Segmentation	HMM-based approach	HMM-based approach	Mean shift with multiple-bandwidth partitioning	HMM-based approach

AS-GENSENG, and several state-of-the-art methods to these data, including GENSENG (Szatkiewicz et al., 2013), CNVnator (Abyzov et al., 2011), and ERDS (Heinzen et al., 2012), using the recommended parameter setups and QC filters. For example, with CNVnator (Abyzov et al., 2011) we used q_0 filter which filters out any predictions that have >50% reads with zero-valued MAPQ (i.e. reads with multiple mapping locations). With ERDS (Heinzen et al., 2012) we removed deletions that are <10kbps and do not have supporting read-pairs. The methodological differences between AS-GENSENG, GENSENG, CNVnator, and ERDS are detailed in Table 3.2. AS-GENSENG differs from existing methods mainly in its incorporation of AS information and simultaneous bias correction.

We simulated two sets of data to assess performance. In the first simulation, we generated 100 TReC plus ASReC datasets. Using chromosome 1 from HapMap sample NA12891 as a template (which provides TReC, covariates, and ASReC), we implanted 200 CNVs (around 60% deletions and 40% duplications, median size = 3,000 bp) by modifying the original TReC. We assigned a randomly generated copy number range of 3-6 for duplications and 0-1 for deletions (all other windows were assigned copy number two). We passed the covariate matrix (columns were the assigned copy number, mappability score, and GC content respectively; rows were each sliding window) and coefficient vector (all initialized as 1) to the garsim function from R/gsarima to simulate TReC for each window. We applied the NB distribution with the log-link function for the garsim model, where the autoregressive parameter was set to 0.6 (because we used overlapping windows); the zero-correction parameter was set to "zq1" and the inverse of the overdispersion

parameter was set to 0.01. As a result, TReCs changed over the CNV windows. In the next step we simulated ASCNV. For each simulated CNV, we computed the mean ASReC of all its spanned windows in the template. If the mean ASReC was larger than a threshold (>10), we called it AS informative. We declared AS informative CNVs as ASCNV and simulated their allelic configurations (i.e., the number of copies from alleles A and B). We then simulated ASReC ($o^{(A)}$ and $o^{(as)}$) based on allelic configurations. The $o^{(as)}$ for each window was obtained from the template; and the $o^{(A)}$ for each window was computed as follows: 1) If $o^{(as)} \leq 10$ or the window was not in a CNV, $o^{(A)} = 0.5 * o^{(as)}$; 2) Otherwise, we defined p as the proportion from allele A from the allelic configuration, and $o^{(A)} = p * o^{(as)}$.

In the second simulation, we mimicked the sequencing experiment by generating paired end reads from a hypothetical chromosome. To simulate reads, we first simulated one pair of chromosomes using human chromosome 1 as the template and implanted 200 CNVs by modifying the sequence (around 60% deletions and 40% duplications, median size = 3000bps). We specified the allelic configuration for each CNV so that the modification was applied to the proper side of the chromosome (e.g. for a copy number four duplication, if the allelic configuration was ABBB, we duplicated the sequence only from the B allele two more times; if its allelic configuration was AABB, we duplicated the sequences from both alleles once). After creating the hypothetical chromosome, we applied the sequencing simulator, wgsim, as implemented in SAMTools (Li et al., 2009) to generate 100bps paired-end short reads. We used the default wgsim values. In total, 50 millions read pairs were generated and yielded on average 40X coverage. We extracted allele specific reads using NA12891 heterozygous SNPs. Finally, we used BWA (Li and Durbin, 2009) to align the reads back to unmodified reference to obtain TReC and ASReC.

As detailed in the section Data preparation, the empirical datasets used in our evaluation consisted of high-coverage WGS data from two HapMap individuals (NA12891, NA12892) sequenced as a part of the 1000GP (Mills et al., 2011; Abecasis et al., 2012; Handsaker et al., 2011).

To evaluate sensitivity and false discovery rate (FDR), we focused on autosomal CNVs and intersected the predicted CNVs of different methods with the known CNVs. The known CNVs

(Figure 2.7) are either the simulated ground-truth CNVs for the simulated data or the 1000GP-released high-confidence deletions for the empirical data. Sensitivity was calculated as the proportion of known CNVs overlapped by predicted CNVs with the correct CNV type. Following 1000GP, we defined CNV type as deletions (integer copy number 0 or 1) or duplications (integer copy number ≥ 3). Sensitivity for detecting duplications in the two HapMap individuals was not evaluated because of the lack of high-confidence duplications (Mills et al., 2011; Abecasis et al., 2012). The FDR for the simulation data was calculated as the proportion of predicted CNVs not overlapped with the known CNVs. Because the true negatives for the two HapMap individuals are not known, we used the total number of base pairs and the total number of calls as a surrogate measure for specificity. A 50% reciprocal overlap was used as the overlapping criterion in all comparisons.

To evaluate the AS information, we reported the ASCN set. In the simulation, we had the ground truth ASCN set (ASReC > 10). Thus we reported the sensitivity and FDR based on the ground truth ASCN set, and compared with the values on the entire set. With the empirical data, we reported the number of detected ASCNs.

CNV detection performance depends on sequencing coverage, especially for read-depth-based methods. In the comparative analyses conducted in this study, both simulated (40x) and the 1000GP data (>30x) had high coverage. Thus we carried out a computational experiment in order to identify the lower bound on sequencing coverage that AS-GENSENG can handle. In this experiment, we first used Picards DownSampleSam.jar tool to down-sample the high coverage data to varying coverage of 30x, 20x, 10x, and 5x; and next applied AS-GENSENG to each resulting dataset and evaluated the performance using the same metrics.

3.2.6 Performance Evaluation: WES data

We used both simulation and empirical data to assess the performance of AS-GENSENG in detecting CNVs and ASCNVs in WES data. For the purpose of methods comparison, we applied both AS-GENSENG and several state-of-the-art WES CNV detection methods to these data, including Conifer (Krumm et al., 2012), XHMM (Fromer et al., 2012) and ExomeDepth (Plagnol

Table 3.3: Methodologies comparisons among WES methods

	AS-GENSENG	XHMM	Conifer	ExomeDepth
Raw Data	Total Read Count (TReC) Allele Specific Read Count (ASReC)	Depth of Coverage (DOC)	Read Count (RC)	Read Count (RC)
Normalization	Compute the ratio value for each exon target as the TReC of the target to the sum of TReCs of all targets for each sample. Apply ASReC to find a subset of samples which do not have CNV at the given exon. Median ratio value of the no CNV samples group as the baseline ratio at the exon. Multiplies baseline ratio with the sum of TReCs of a sample to compute the expected TReC of the sample at a given exon.	Apply PCA to remove high-variance noise	Convert RC to RPKM. Compute mean RPKM value of all samples as the baseline of a given exon (not accurate when common CNV exists). Convert RPKM to z-score based on the mean value, and the standard deviation of RPKM values of all samples.	Build an optimized reference set using a beta-binomial distribution. Compare the read count with the corresponding read count in the built reference set.
CNV calling	Hidden Markov Model	Hidden Markov Model	Segmentation based on fixed threshold based on standard deviation.	Hidden Markov Model
Absolute copy number	Directly model absolute copy number as the hidden states.	No absolute copy number calls.	Required external sources of CNV with absolute copy number.	No absolute copy number calls.

et al., 2012). We used the recommended parameters and QC filters for each method. For example, with Conifer, we removed five SVD components for detecting common CNVs (5%, 10%, and 20% frequency) and 10 components for detecting rare CNVs (1% frequency) (Krumm et al., 2012). With XHMM, we used the default value 30 for CNV quality threshold (Fromer et al., 2012). Methodological differences are detailed in Table 3.3. The primary novel aspect of AS-GENSENG is its use of allele specific information in the modelling and its ability to detect ASCN.

We developed a pipeline to simulate the TReC and ASReC for 100 samples (Figure 3.8). It consists of 11 steps summarized here. First, we selected chromosome 11 of one HapMap sample (HG00264) as template. From this template we obtained the exome-capturing target, its TReC, and the sum of the TReCs of all targets. We retained only targets in copy-number-two regions, which were extracted from 1000GP-released CNV analysis results. The next step was to simulate TReC. For each target in the template, we calculated the ratio between its TReC and the sum TReCs of all targets; this serves as the expected ratio of the target. We first simulated the sum of the TReCs of all the targets, to provide a sample that followed a normal distribution with the mean equal to the sum of the template TReCs. We then simulated the expected TReCs, target by target. To simulate the expected TReC for one target, we multiplied the template expected ratio by the sum TReCs of the sample. After simulating the expected TReCs for all targets, we use the garsim function to simulate the overdispersed TReCs. We passed the covariates matrix (using the expected TReCs at each target as the mean value) and coefficient vectors (all 1 for targets) into the garsim function, setting the distribution family as negative-binomial distribution, and the overdispersion parameter as 0.05. The

result gave TReCs at each target for one sample. We repeated this procedure 100 times to obtain the TReCs for 100 samples. The next step was to simulate CNV with different population frequencies. To simulate a CNV, we needed to choose the targets and the samples affected by this CNV. For each CNV, we randomly chose 13 consecutive targets and reduced the expected TReC (e.g., multiplying by 0.5 for state 1) or amplified the expected TReC (e.g., multiplying by two for state four) at these targets. To simulate samples with altered copy number in a population, we defined a variable called CNV frequency that identified the proportion of samples (among the 100 simulated samples) with copy number alterations. The CNV frequency ranges used in this study were 1%, 5%, 10% and 20%, to simulate both rare CNV and common CNV. For one CNV, only the TReCs of the affected individuals were altered. We simulated different numbers of CNVs for different CNV frequencies, because when the CNV frequency is low (1% or 5%), a simulated CNV would affect only a small number of samples, which would cause the average number of CNVs in each sample to be extremely low. This in turn made it difficult to give a reasonably large denominator in the calculations of sensitivity and FDR. Therefore, we simulated 1000 deletions/1000 duplications with CNV frequencies of 1% and 5% and 200 deletions/200 duplications with frequencies of 10% and 20%. The last step was to simulate ASReC. We first simulated ASReC ($o^{(as)}$),following a binomial distribution with the mean value being 5% of the RD in the corresponding target. We then simulated the allelic configuration according to the copy number of the target (e.g., AAB or ABB for copy number 3) and obtained the proportion p of allele A (e.g., 0.33 for ABB). Finally we simulated the ASReC of allele A, which follows a beta-binomial distribution with the mean value of p multiplying the corresponding ASReCs and the overdispersion being 0.1.

As detailed in the section Data preparation, the empirical data used in our evaluation consisted of WES data from 324 HapMap individuals sequenced as a part of the 1000GP (Mills et al., 2011; Abecasis et al., 2012). High-confidence exonic-deletions for these samples were obtained by intersecting the 1000GP-released high-confidence genome-wide deletions with exon targets (>1bp overlap) (Mills et al., 2011; Abecasis et al., 2012; Tan et al., 2014).

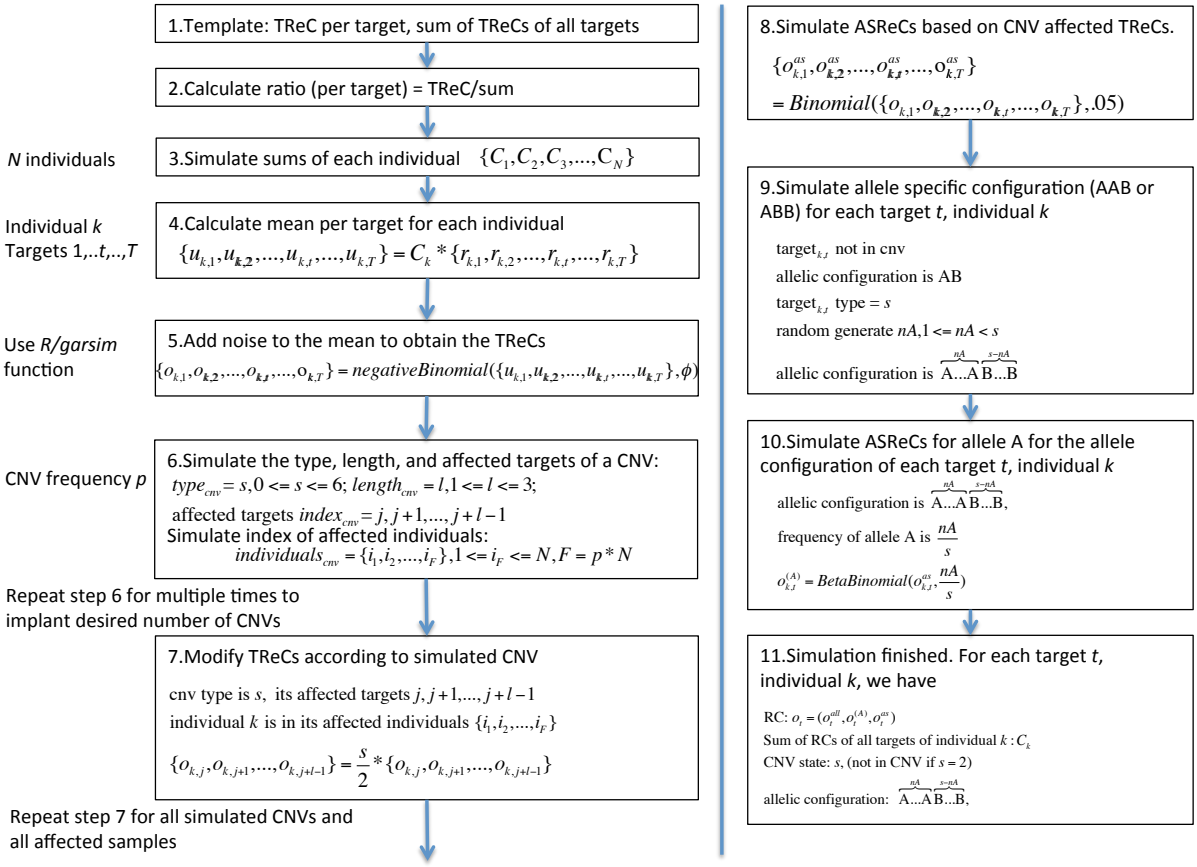


Figure 3.8: Whole-exome sequencing-simulation flowchart

To evaluate sensitivity and FDR, we focused on autosomal CNVs and intersected the predicted CNVs of different methods with the known CNVs (i.e. the ground-truth CNVs used in simulation or the high-confidence exonic-deletions (see Data Preparation) for the empirical data). Sensitivity was calculated as the proportion of known CNVs overlapped by predicted CNVs with the correct CNV type (Figure 2.7). For the empirical data, sensitivity for detecting duplications was not evaluated due to the lack of high-confidence duplications in the literature (Mills et al., 2011; Abecasis et al., 2012). For the simulated data, FDR was calculated as the proportion of predicted CNVs not overlapped with the known CNVs. For the empirical data, FDR was evaluated using the SuperArray Validation (SAV) developed by the 1000GP (Mills et al., 2011; Abecasis et al., 2012). The SuperArray integrated available intensity data for HapMap samples from three array platforms (Affymetrix 6.0, Illumina 1M, and a custom Nimblegen aCGH array with 4,938,838 probes) into a

high-density virtual array. A non-parametric testing procedure is developed to calibrate predicted CNVs using SuperArray. The rule of thumb of the procedure is that the intensity data of samples with lower underlying copy number tend to be lower than samples with higher underlying copy number. A Wilcoxon Rank Sum (WRS) test was carried out on each predicted CNV region (where at least one sample made a CNV call. To ensure statistics power of the WRS tests, we required that at least 15 deletion or 13 duplication ranks existed in a predicted CNV region. FDR was estimated as two times the fraction of predicted regions with p-value $>$0.5.

To evaluate the AS information, we reported the ASCN set. In the simulation, we had the ground truth ASCN set (ASReC $>$10), so we reported the sensitivity and FDR on the ground truth ASCN set. In the empirical data, we reported the number of detected ASCNs.

3.2.7 CNV validation using NanoString technology

In order to validate randomly selected deletions and duplications, we utilized an independent methodology, NanoString nCounter, a proven and high-throughput method for CNV verification (Geiss et al., 2008; Sailani et al., 2013; Iskow et al., 2012; Ruderfer et al., 2013; Brahmachary et al., 2014). We focused on validating AS-GENSENGs ability to detect CNVs from WES data and NA12272 was randomly chosen from the 324 HapMap samples for which WES data was analysed. Our first goal was to compare AS-GENSENG calls in sample NA12272 with the relative copy number estimated by NanoString (following the analysis method in (Brahmachary et al., 2014)). It is important to note that, for each probe, NanoString requires samples known to be copy number two, so we relied on the absolute copy number reported in Conrad et. al. (Conrad et al., 2010) to calibrate the NanoString calls (i.e. indicate which samples have copy number two). In addition, we would separate the calls with overlapping SNPs from the calls without overlapping SNPs to study the effect of using SNP information in CNV detection.

Our second goal was to compare AS-GENSENG calls in sample NA12272 with two other members of the trio (paternal: NA12272, maternal: NA12273, and child: NA10837) in order to identify Mendelian inconsistencies. DNA for each of these samples was acquired from the Coriell repository (`http://ccr.coriell.org`) and used as input for the NanoString nCounter CNV

Table 3.4: Performance comparison on WGS-read-count-simulation data

Simulation	Average Implanted CNVs		Average Number of True Discoveries (Sensitivity)		Average Reported CNVs		Average Number of False Discoveries (FDR)	
	All	w AS	All	w AS	All	w AS	All	w AS
Deletions	121.7	11.8	112.6 (92%)	11.5 (98%)	114	11.6	0.8 (0.7%)	0.1 (0.7%)
Duplications	78.3	15.5	67.5 (86%)	13.9 (90%)	97.8	13.9	29.9 (30.6%)	0 (0.0%)
Overall CNV	200	27.3	180 (90%)	25.3 (93%)	211.7	25.5	30.7 (14.5%)	0.1 (0.3%)

assay, according to the manufacturers instructions. In short, 600 ng of genomic DNA was fragmented to $\sim$500 bp by digestion with AluI and subject to a multiplex hybridization reaction involving all probes. We designed a custom NanoString probe set (using NanoStrings nDesign Gateway software) targeting 11 deletions and 14 duplications predicted by AS-GENSENG, with each locus targeted by a single custom probe. The custom probes were 70-100bp in length, each was placed in the middle of a targeted CNV and all satisfied the internal design parameters used by NanoString, such as good GC-content and not-overlapping segmental duplication or repetitive elements. The probe set also included eight negative control probes that target artificial sequences, and ten normalization probes that target autosomal loci that are invariant in copy number. Data analysis was conducted as in (Brahmachary et al., 2014).

3.3 Results

3.3.1 CNV detection in whole-genome sequencing-simulation data

In order to assess the performance of our method for predicting CNVs from WGS data, we applied AS-GENSENG to two sets of simulated data. We first conducted 100 simulations of TReC and ASReC affected by implanted CNVs. We expected both the TReC and ASReC within the CNV-implanted windows to be affected. Since not every CNV region has enough allele-specific reads to provide informative ASReC, in this work, we defined a CNV region informative for ASCN (i.e., ASCNV) if the ASReC was $>$10. Figure 3.3 shows an examples of simulation, including simulated copy number, TReC at each simulated genomic region, simulated ASCNs, and covariates (GC content and mappability). For each simulation, we implanted 200 CNVs (122 deletions and 78 duplications on average); 27 were ASCNVs (12 deletions and 15 duplications) on average. We estimated sensitivity and false-discovery rate (FDR) by intersecting the AS-GENSENG-predicted CNVs with the implanted CNVs (using $\geq$50% reciprocal overlapping as the criterion; illustrations

Table 3.5: Performance comparison on WGS-sequencing-read-simulation data

Simulation	Implanted CNVs		Number of True Discoveries (Sensitivity)		Reported CNVs		Number of False Discoveries (FDR)	
	All	w AS	All	w AS	All	w AS	All	w AS
Deletions	119	15	117 (98%)	15 (100%)	119	15	2 (2%)	0 (0%)
Duplications	81	47	76 (94%)	46 (98%)	88	34	12 (14%)	2 (6%)
Overall CNV	200	62	193 (97%)	61 (98%)	207	49	14 (7%)	2 (4%)

Table 3.6: Performance comparison with WGS methods on WGS-simulation data

CNV type	Average Implanted CNVs	#True Discoveries[3] (Sensitivity)			Predicted CNVs			#False Discoveries[3] (FDR)		
CNV frequency		AS-GENSENG	ERDS	CNVnator	AS-GENSENG	ERDS	CNVnator	AS-GENSENG	ERDS	CNVnator
Deletion	119	117 (98%)	113 (95%)	107 (90%)	119	113	138	2 (2%)	0 (0%)	31 (22%)
Duplication	81	76 (94%)	78 (96%)	75 (93%)	88	156	76	12(14%)	78 (50%)	1 (1%)
Overall	200	193 (97%)	191 (96%)	189 (91%)	207	269	214	14 (7%)	78 (30%)	32 (15%)

of sensitivity and FDR calculation are shown in Figure 2.7). Results are detailed in Table 3.4 and are summarized below. On average, AS-GENSENG predicted 180 ground-truth CNVs (90% sensitivity) from each simulated dataset, including 113 deletions (92% sensitivity) and 68 duplications (86% sensitivity). Regarding ASCNVs, AS-GENSENG predicted 25 ground-truth ASCNVs (93% sensitivity), including 11 AS deletions (98% sensitivity) and 14 AS duplications (90% sensitivity). Therefore, we observe slightly higher sensitivity for detecting AS events and, furthermore, the FDR for ASCNVs is 0.3%, much lower than that for standard CNVs (14.5%). This lower FDR results from the fact that read-count signals alone are vague for differentiating between copy number two and copy numbers three or one. Without ASReC, the algorithm could make false-positive calls. However, the difference in allele-specific proportion is much clearer between copy number two and copy numbers three or one (0.5 in copy number two compared to 0.33 or 0.67 in copy number three, and 0.99 or 0.01 in copy number one). As a result, the FDR of ASCNVs is much lower.

We next simulated sequencing reads affected by implanted CNVs. We simulated 200 CNVs (119 deletions and 81 duplications); 62 were ASCNVs (15 deletions and 47 duplications) and estimated sensitivity and FDR using the same criterion. We first compared AS-GENSENGs performance for standard CNVs versus ASCNVs, and then compared AS-GENSENG to CNVnator and ERDS for the ability to detect standard CNVs. Results are detailed in Tables 3.5-3.6 and

[3] $\geq$50% overlap

summarized below. For standard CNVs, AS-GENSENG predicted 193 ground-truth CNVs resulting in 97% sensitivity (117 deletions with 98% sensitivity and 76 duplications with 94% sensitivity). For ASCNVs, AS-GENSENG predicted 61 ground-truth AS-CNVs resulting in 98% sensitivity (15 AS deletions with 100% sensitivity, and 46 AS duplications with 98% sensitivity), slightly higher than that for standard CNVs. Further, AS-GENSENGs FDR for ASCNVs is 4%, lower than that for standard CNVs (7%). Compared to other methods on the ability to detect standard CNVs, AS-GENSENG had the highest sensitivity and lowest FDR (sensitivity 6% higher than CNVnator and 1% higher than ERDS; FDR 8% lower than CNVnator and 23% lower than ERDS).

In summary, simulation results suggested that incorporating ASReC improves the sensitivity and specificity of CNV detection.

3.3.2 CNV detection in whole-genome-sequencing real data

To further evaluate the performance of our method for WGS data, we analysed 1000GP (Mills et al., 2011; Abecasis et al., 2012) data. We applied AS-GENSENG, GENSENG (Szatkiewicz et al., 2013), CNVnator (Abyzov et al., 2011), and ERDS (Heinzen et al., 2012) to the high-coverage WGS data for samples NA12891 and NA12892 and compared the predicted CNVs to a high-confidence, published dataset available for these samples (Mills et al., 2011; Abecasis et al., 2012; Handsaker et al., 2011) (2200 deletions in NA12891 and 2055 deletions in NA12892 but no high-confidence duplications available). AS-GENSENG differs from existing methods mainly in its incorporation of AS information for both deletions and duplications (in comparison to GENSENG, CNVnator, and ERDS) and its simultaneous bias correction (in comparison to CNVnator and ERDS). The methodological differences between these methods are detailed in Table 3.2.

First, we compared AS-GENSENG to GENSENG and CNVnator, both relying only on TReC for CNV detection. As shown in Table 3.7a, the sensitivity for detecting deletions by AS-GENSENG was 56% for NA12891 and 53% for NA12892, which is higher than GENSENG (50% for NA12891, 49% for NA12892) and CNVnator (37% for NA12891 and 34% for NA12892). We also found several examples of high-confidence deletions that were missed by GENSENG but

Table 3.7: Performance assessment based on WGS data of two HapMap samples

a Comparison with GENSENG and CNVnator

	#Deletions (total Mbps)			#True Discovery Deletions/ #HC Deletions (Sensitivity)		
Genome	AS-GENSENG	GENSENG	CNVnator	AS-GENSENG	GENSENG	CNVnator
NA12891	2302 (20.4)	4765 (88.1)	2656 (131.3)	1222/2200 (0.56)	1091/2200 (0.50)	815/2200 (0.37)
NA12892	2347 (24.3)	4295 (45.0)	2268 (128.0)	1079/2055 (0.53)	1006/2055 (0.49)	698/2055 (0.34)

b Comparison with ERDS

	# Deletions (total Mbps)		#True Discovery Deletions/ #High Confidence Deletions (Sensitivity)							
Genome	AS-GENSENG	ERDS	AS-GENSENG				ERDS			
			<1k bps	1k-10k bps	>10k bps	total	<1k bps	1k-10k bps	>10k bps	total
NA12891	2302 (20.4)	1911 (19.6)	223/1132 (0.2)	803/863 (0.93)	196/205 (0.97)	1222/2200 (0.56)	483/1132 (0.43)	608/863 (0.70)	145/205 (0.71)	1236/2200 (0.56)
NA12892	2347 (24.3)	1712 (17.5)	217/1121 (0.19)	681/750 (0.91)	181/184 (0.98)	1079/2055 (0.53)	457/1121 (0.41)	521/750 (0.69)	140/184 (0.76)	1118/2055 (0.54)

were recovered by AS-GENSENG (Figures 3.4,3.5). Due to the relatively high level of noise in the TReC of these relatively small deletions (<10 windows), the TReC signal by itself does not provide enough evidence for GENSENG to call deletions. However, the imbalance of ASReC signals in these examples strongly implies underlying CNVs. Thus, by incorporating ASReC with TReC, AS-GENSENG successfully recovered these deletions.

We then compared the specificity of various methods. Because the high-confidence CNV dataset does not provide information on the true negatives for assessing specificity, we used the volume (i.e., the total number and total base pairs) of the predicted CNVs as a surrogate measurement of specificity. As shown in Table 3.7a, the volume of AS-GENSENG is much smaller than GENSENG and CNVnator, suggesting improved specificity. Second, we compared AS-GENSENG to another integrated method, ERDS. ERDS incorporates the rate of heterozygous SNPs in detecting deletions and further refine the smallest deletion calls (<10kbps) using read-pair information; but ERDS relies only on TReC in detecting duplications. Thus in our sensitivity evaluation, we stratified the comparative analysis by the size of the high-confidence deletions in three categories (<1kb,1kb-10kb and >10kb). Finally, we applied AS-GENSENG, CNVnator and ERDS to the WGS data from a HapMap trio (NA12891, NA12892, NA12878) and computed the rate of Mendelian inconsistencies as a measure of specificity. By intersecting CNV calls in the child (NA12878) with CNV calls in the parents, we found that AS-GENSENG had the lowest Mendelian error rate (25%) among all CNVs predicted in the child (23% for deletions and 28% for

duplication), whereas CNVnator had 48% Mendelian errors (47% for deletions and 52% for duplications) and ERDS had 55% Mendelian errors (48% for deletions and 57% for duplications).

As shown in Table 3.7b, in the >10kb category when both ERDS and AS-GENSENG incorporate SNP information with TReC, AS-GENSENG achieved 26% higher sensitivity in NA12891 (97% vs. 71%) and 22% higher sensitivity in NA12892 (98% vs. 76%). In the 1kb-10kb category even after ERDS applied read-pair information for call refinement, AS-GENSENG achieved 23% higher sensitivity in NA12891 (93% vs. 70%) and 22% higher sensitivity in NA12892 (91% vs. 69%). Only in the <1kb category AS-GENSENG had lower sensitivity than ERDS, which can be attributed to two factors: (1) AS-GENSENG can only detect CNVs of twice or more of the window size (i. e.> 600bp), whereas ERDS does not have this limitation; (2) Within the category of CNVs > 600bp and smaller than 1kb, ERDS has advantage by additionally using read-pair information. We then examined the volume of CNV calls as a surrogate measure of specificity. AS-GENSENG predicted slightly higher number of deletions than ERDS (2303 vs. 1911 in NA12891 and 2347 vs. 1712 in NA12892), suggesting comparable specificity; but predicted a much smaller number of duplications than ERDS (723 vs. 3404 in NA12891 and 581 vs. 3432 in NA12892), suggesting improved specificity.

In summary, when applied to high-coverage WGS data, AS-GENSENG outperforms existing methods for detecting deletions that are >1Kb. It gives the best sensitivity (∼5% higher than GENSENG, ∼20% higher than CNVnator, and more than 20% higher than ERDS) and among the best specificity (only slight larger than ERDS in the deletion calls). These results suggest that incorporating AS information improves the accuracy of CNV detection. Further, in regard to ASCNVs, AS-GENSENG is the only method that can predict ASCN call from WGS data in germline DNA samples. In this experiment, AS-GENSENG predicted 576 AS deletions and 205 AS duplications in NA12891, 664 AS deletions and 173 AS duplications in NA12892.

As expected, we find that the higher the sequencing coverage, the better the performance for AS-GENSENG to detect CNVs. The lowest bound of sequencing coverage that AS-GENSENG still

[4] ≥ 50%bp overlap

Table 3.8: Performance comparison on WGS simulation down-sampled data

Parameter	Deletion			Duplication		
Coverage	#True Discoveries[4]/#Implanted CNV	Sensitivity	#False Discoveries/#Discoveries	FDR	#True Discoveries[4] /#Implanted CNV	Sensitivity	#False Discoveries/#Discoveries	FDR
5x	57/119	48%	0/57	0%	52/81	64%	0/52	0%
10x	90/119	77%	2/92	2%	64/81	79%	3/67	4%
20x	111/119	93%	3/114	3%	74/81	92%	5/79	6%
30x	114/119	96%	3/117	3%	75/81	93%	8/83	10%
40x	117/119	98%	2/119	2%	76/81	94%	12/88	14%

Table 3.9: Performance comparison on WGS simulation down-sampled data

Parameter[5]	NA12891			NA12892		
Coverage	#True Discoveries[6]/#High Confidence Deletions	Sensitivity	#AS-GENSENG deletion calls/Spanned regions in Mbps	#True Discoveries[6] /#High Confidence Deletions	Sensitivity	#AS-GENSENG deletion calls/Spanned regions in Mbps
5x	768/2200	35%	684/6.8	652/2055	32%	601/5%
10x	1018/2200	46%	1403/8.4	880/2055	43%	1360/7.3
20x	1178/2200	54%	1817/13.7	1044/2055	51%	1897/11.3
25x	-	-	-	1079/2055	53%	2347/24.3
30x	1222/2200	56%	2302/20.4	-	-	-

achieves a reasonable sensitivity is 10x (Tables 3.8-3.9). At a low coverage of 5x, AS-GENSENGs sensitivity is remarkably reduced (i.e. 29% reduction in detecting deletions and 15% reduction in detecting duplications in simulation study, and 11% reduction in deletions for 1000GP WGS data).

3.3.3 CNV detection in whole-exome sequencing-simulation data

In order to calibrate the performance of our method for WES data, we applied AS-GENSENG to simulated datasets and evaluated the sensitivity and FDR by comparing the predicted CNVs with the implanted ground-truth CNVs (Table 3.10). In particular, we evaluated

Table 3.10: Performance comparison on WES-simulation data

Simulation Parameter	Average Implanted CNVs				Average #True Discoveries[7] (Sensitivity)				Average Predicted CNVs				Average #False Discoveries[7] (FDR)			
CNV frequency	Deletions		Duplications		Deletions		Duplications		Deletions		Duplications		Deletions		Duplications	
	All	w AS[8]	All	w AS	All	w AS	All	w AS	All	w AS	All	w AS	All	w AS	All	w AS
0.01	10	2.3	10	8.1	8.9 (89%)	2.1 (95%)	9.1 (91%)	7.5 (93%)	9.0	2.0	9.5	7.5	0.01 (0%)	0 (0%)	0.0 (0%)	0 (0%)
0.05	50	10.4	50	40.3	40.3 (81%)	9.5 (91%)	45.6 (91%)	37.7 (93%)	44.0	9.0	47.2	36.6	0.03 (0%)	0.01 (0%)	0.04 (0%)	0.01 (0%)
0.1	20	3.7	20	15.6	16.4 (82%)	3.4 (91%)	17.9 (89%)	14.6 (93%)	17.6	3.2	18.7	14.2	0.01 (0%)	0 (0%)	0.01 (0%)	0 (0%)
0.2	40	8.5	40	30.6	34 (85%)	7.8 (92%)	35.9 (90%)	28.1 (92%)	36.0	7.5	38.7	28.3	0.44 (1.3%)	0 (0%)	0.03 (0%)	0.01 (0%)

[5]The coverage of NA12891 sequencing data provided by 1000GP is 30x. We down-sampled sequencing reads from the 1000GP sequencing data with the target coverage from 5x, 10x, to 20x. The coverage of NA12892 sequencing data provided by 1000GP is 25x. We down-sampled sequencing reads from the 1000GP sequencing data with the target coverage from 5x, 10x, to 20x.

[6]$\geq$ 50%bp overlap

[7]>1bp overlap

[8]mean ASReC>10

Table 3.11: Performance comparison with WES methods on WES-simulation data

Simulation Parameter	Average Implanted CNVs	Average #True Discoveries[9] (Sensitivity)				Average Predicted CNVs				Average #False Discoveries[9] (FDR)			
CNV frequency		AS-GENSENG	Exome-Depth	Conifer	XHMM	AS-GENSENG	Exome-Depth	Conifer	XHMM	AS-GENSENG	Exome-Depth	Conifer	XHMM
1%	20	17.91 (89.6%)	16.48 (82.4%)	10.66 (53.3%)	17.53 (87.7%)	18.54	16.41	10.5	17.63	0.01 (0%)	0.01 (0%)	0.04 (0.4%)	0.34 (1.9%)
5%	100	85.86 (85.9%)	35.13 (35.13)	29.63 (29.63%)	5.99 (5.99%)	91.32	33.28	27.95	5.78	0.07(0%)	0.00 (0%)	0.02 (0%)	0.07 (1.2%)
10%	40	34.33 (85.8%)	24.36 (60.9%)	4.99 (12.5%)	0.82 (2.1%)	36.28	24.36	5.49	0.83	0.02 (0%)	0.00 (0%)	0.08 (1.5%)	0.01 (1.2%)
20%	80	69.89 (87.4%)	27.36 (34.2%)	1.04 (1.3%)	0.12 (0.2%)	73.67	27.38	1.57	0.13	0.47 (1%)	0.02 (0%)	0.37 (23.5%)	0.01 (7.7%)

AS-GENSENGs ability to detect CNVs at varying allele frequencies. Following the criterion of rare CNVs ($<$5% in the population (39)), we simulated both rare (1%) and common CNVs (5%, 10%, and 20%). First, we calculated sensitivities (using $>$1bps reciprocal-overlap) for the entire set of implanted CNVs or ASCNVs. The sensitivities ranged from 81% to 91% for various CNV frequency settings and there was no remarkable difference in sensitivity between the rare CNV and common CNV sets. For example, the respective sensitivities are 0.89 for deletions and 0.91 for duplications on the 1% CNV frequency set, and 0.85 for deletions and 0.90 for duplications on the 20% CNV frequency set. Regarding ASCNVs, all sensitivities were $>$90% and better than the corresponding regular CNV values. For example, with a CNV frequency of 5%, we observed a 10% improvement for deletions and 2% improvement for duplications. Second, we evaluated FDR and found that AS-GENSENG had very low FDR for both CNV and ASCNV (most $<$1%). These results suggest that, when applied to WES data, AS-GENSENG can robustly detect both rare and common CNVs at varying frequencies, and that incorporating ASReC improves the accuracy for CNV detection.

Next, using the same simulated WES datasets, we compared AS-GENSENG with three state-of-the-art methods, XHMM (Fromer et al., 2012), Conifer (Krumm et al., 2012) and ExomeDepth (Plagnol et al., 2012) (see Table 3.3 for detailed method comparison). XHMM and Conifer use an SVD-based normalization (Fromer et al., 2012; Krumm et al., 2012). ExomeDepth uses a reference-based normalization with an optimized reference-set (Plagnol et al., 2012). AS-GENSENG uses a reference-based normalization and its novelty is its explicit use of ASReC to

[9] $>$1bp overlap

identify the correct reference group of copy-number-two, which is critical for data normalization and the detection of common CNVs of unknown frequencies. The results of the comparative analysis are summarized in Table 3.11. We find that the sensitivity of AS-GENSENG is higher than XHMM, Conifer and ExomeDepth for all CNV frequencies, especially in detecting common CNV (AS-GENSENG sensitivity is 89.6 % for CNV frequency 1% while XHMM is 87.7% Conifer is 53.3% and ExomeDepth sensitivity is 82.4%; AS-GENENG sensitivities are higher than 80% for CNV frequency $\geq$ 5% while XHMM sensitivities are less than 6, Conifer sensitivies are less than 30% and ExomeDepth sensitivities are less than 61%). While AS-GENSENG demonstrated consistently good sensitivity across frequency categories, the sensitivities of XHMM, Conifer and ExomeDepth decrease in common CNVs. For the most common CNVs (frequency =20%), AS-GENSENG was $>$ 100X more sensitive than XHMM, 60X more sensitive than Conifer and 2.5X more sensitive than ExomeDepth (AS-GENSENG sensitivity is 87.4%; XHMM is 0.2%, Conifer is 1.3%; ExomeDepth is 34.2%).

Presumably, at higher CNV frequencies, CNV signals may have stronger contributions to the very variance components that are excluded by the SVD method with an arbitrary threshold (Tan et al., 2014); and TReC may not identify the true reference copy number two set. We find that the FDR of AS-GENSENG is $<$1% for all settings, the FDR of XHMM is $<$ 2% for most settings, the FDR of Conifer is $<$2% for most settings, and the FDR of ExomeDepth is $<$1% for all settings, suggesting similar, high specificity. In summary, these results suggest that reference-based normalization combined with assumption-free identification of the copy number two reference, such as using ASReC as implemented in AS-GENSENG, is critical for robust detection of common CNVs at varying frequencies that cannot be known a *priori*.

3.3.4 CNV detection in real whole-exome-sequencing data

To further evaluate the performance of CNV detection in WES data, we applied AS-GENSENG, Conifer (Krumm et al., 2012), XHMM (Fromer et al., 2012), and ExomeDepth (Plagnol et al., 2012) to the WES data of 324 HapMap samples (Table 3.12). The total numbers of CNVs called from AS-GENSENG, Conifer, and XHMM are comparable. AS-GENSENG predicted

Table 3.12: Performance assessment based on WES-emprical data

Method	#Reported Deletions	Estimated FDR[10] for Deletions	#True Discovery Deletions[11]/#1000GP Released Deletions (Estimated Sensitivity)		
			Deletions by frequency in samples		
			<5%	>5%	All
AS-GENSENG	4839	19.2%	90/506 (17.8%)	1503/8686 (17.3%)	1593/9192 (17.3%)
Conifer	2194	12.8%	160/506 (31.6%)	736/8686 (8.5%)	896/9192 (9.7%)
XHMM	3006	44.4%	244/506 (48.2%)	135/8686 (1.6%)	801/9192 (8.7%)
ExomeDepth	74463	95.5%	77/506 (15.2%)	653/8686 (7.5%)	730/9192 (7.9%)

4839 deletions and 2648 duplications in total from the 324 samples, while Conifer predicted 2194 deletions and 3450 duplications, XHMM predicted 3006 deletions and 3660 duplications. ExomeDepth predicted 74463 deletions and 47816 duplications, which is similar to literature results using ExomeDepth or ∼300 CNVs per sample with around two thirds of CNVs as deletions, ((Plagnol et al., 2012), `http://cran.r-project.org/web/packages/ExomeDepth/vignettes/ExomeDepth-vignette.pdf`), To evaluate sensitivity, we compared these call sets with the high-confidence exonic-deletions as described in Methods; and then repeated this analysis separately for rare (<5% frequency) and common (>5% frequency) high-confidence exonic-deletions. Sensitivity for duplications was not evaluated due to the lack of high-confidence duplication call sets in the literature (Mills et al., 2011; Abecasis et al., 2012). Key results of the sensitivity evaluation are summarized below.

First, AS-GENSENG demonstrated the highest overall sensitivity for detecting high-confidence exonic-deletions (7.6% higher than Conifer, 8.6% higher than XHMM, and 9.4% higher than ExomeDepth). Second, the sensitivity estimates of AS-GENSENG are consistent across CNV frequency categories, whereas the sensitivity estimates of the other three methods varied considerably between rare and common deletions. Third, for rare deletions, XHMM had the highest sensitivity; and for common deletions, AS-GENSENG had the highest sensitivity. Fourth, we note that the relatively low sensitivities of all methods observed in our evaluation are not surprising; similar results have been reported in recently published independent studies (Tan et al., 2014; Guo et al., 2013). This may reflect technological differences between WES (from which the

[10] estimated using *SAV*, see Section 3.2.6

[11] >1bp overlap

CNV call sets were generated) and WGS (from which the high-confidence exonic-deletions were obtained). Typically, WGS is more powerful in detecting CNVs and does not suffer from the additional systematic biases introduced in the exome capturing step (Tan et al., 2014).

To evaluate FDR, we followed the SuperArray Validation (SAV) approach developed by 1000GP (Mills et al., 2011; Abecasis et al., 2012) (see Methods for details of SAV). For deletions, there were 334 predicted deletion regions (where at least one sample has deletion call) in the AS-GENENG calls set and 32 of these regions had p-value $>$0.5 based on the Wilcoxon Rank Sum test, which yielded a FDR of 19.2%. Similarly, the FDR was 12.8% in Conifer calls set (3 with p-value $>$0.5 among a total 47 regions), was 44.4% in XHMM calls set (16 with p-value $>$ 0.5 among a total 72 regions), and was 95.5% in ExomeDepth calls set (747 with p-value $>$0.5 among a total 1564 regions). For duplications, there were 169 predicted duplication regions (where at least one sample has duplication call) in the AS-GENSENG calls set and 20 regions had p-value $>$0.5, which yielded a 23.7% FDR. Similarly, the FDR was 50.5% in Conifer calls set (27 with p-value $>$0.5 among a total 107 regions), 14.9% in XHMM calls set (14 with p-value $>$0.5 among a total 188 regions), and 85.8% in ExomeDepth calls set (391 with p-value $>$0.5 among a total 911 regions). With SAV, the FDR is not defined for CNVs in individual samples but rather to CNV regions in the collection of all samples, and therefore we did not evaluate FDR stratified by frequency as we did in the sensitivity comparison.

In summary, on the 324 WES samples evaluated in this study, AS-GENSENG demonstrated the best sensitivity for detecting common deletions and comparable specificity to other state-of-the-art methods. The performance of AS-GENSENG was consistent across CNV frequency categories, which can be attributed to its ability to accurately identify the reference copy number two group two using ASReC and free of assumptions (see Figures 3.6, 3.9, 3.10, 3.11 for examples). As expected, XHMM had the best sensitivity for detecting rare deletions because its PCA normalization and HMM parameters were optimized to detect rare variants, assuming that most variation in read-depth was due to noise. Finally, in regard to ASCN, AS-GENSENG is the only

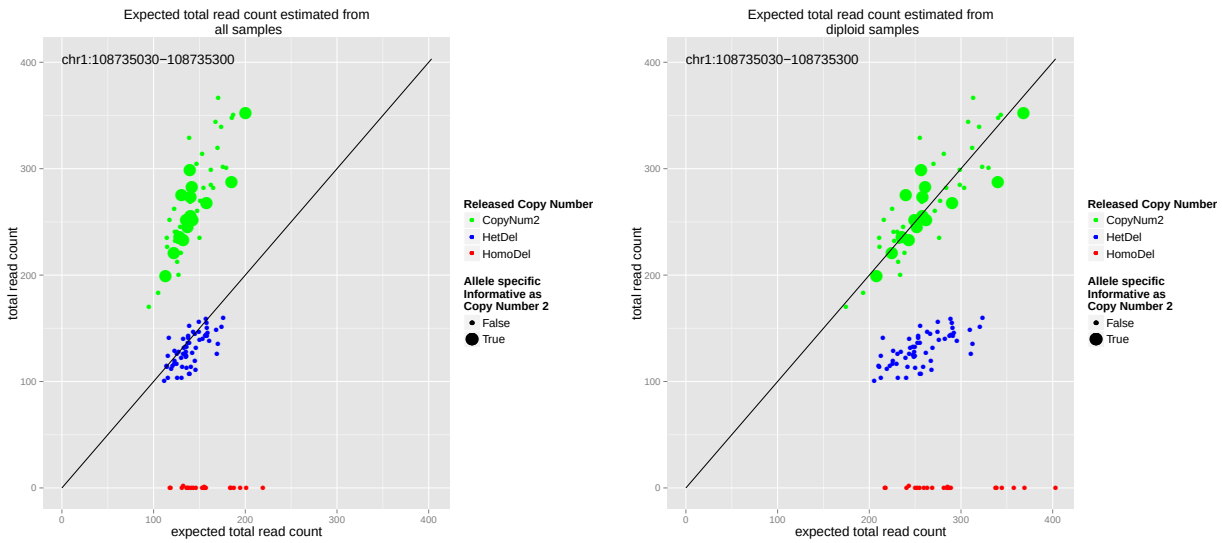


Figure 3.9: An example target that AS-GENSENG better estimates expected RC

method that could predict ASCN from WES data. In this experiment, AS-GENSENG detected 2091 ASCNV (525 AS deletions and 1566 AS duplications) from the 324 HapMap samples.

3.3.5 CNV validation using NanoString technology

We decided to use an independent methodology (NanoString) to validate randomly selected deletions and duplications predicted from WES data. First, we compared AS-GENSENG calls in sample NA12272 with the relative copy number estimated by NanoString (following the method in (Brahmachary et al., 2014)). It is important to note that, for each probe, NanoString requires samples known to be copy number two, so we relied on the absolute copy number reported in Conrad et. al. (Conrad et al., 2010) to calibrate the NanoString calls (i.e. indicate which samples have copy number two). A deletion was validated if its NanoString count is at least 50% smaller than the median of NanoString counts of copy number two samples (Figure 3.12 (a)). Among the 11 randomly selected AS-GENSENG deletions, NanoString identified 9 as true deletions, yielding a validation rate 82% (or 18% false discovery). A duplication was validated if its NanoString count is at least 50% larger than the median of NanoString counts of copy number two samples (Figure 3.12 (b)). Among the 14 randomly selected AS-GENSENG duplications, NanoString identified 12

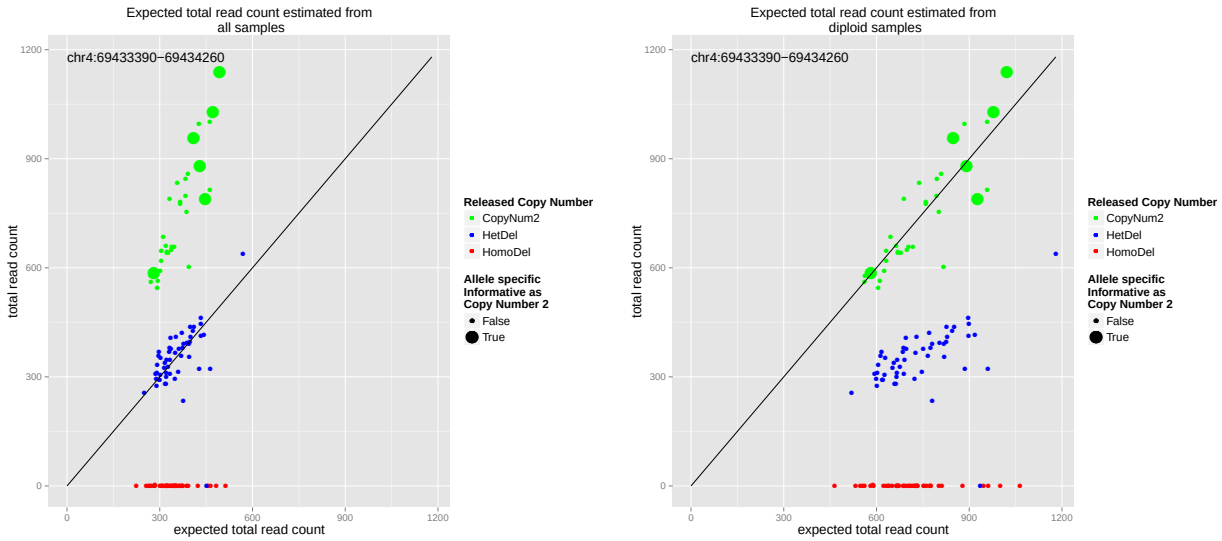


Figure 3.10: An example target that AS-GENSENG better estimates expected RC

as duplications, yielding a validation rate 86% (or 14% false discovery). The results obtained from NanoString validation are similar to the results based on our SuperArray Validation described above.

There are multiple possible explanations for the AS-GENSENG calls that failed to validate. First, it is possible that AS-GENSENG produced false positive calls. Second, it is also possible that false negatives exist in the Conrad et. al. dataset (Conrad et al., 2010) leading to improper calibration to copy number two. Third, and perhaps most likely, the issue could simply be a matter of probe placement, since we tested just one probe per CNV region. We decided to test a limited number of probes per CNV in order to maximize the number of CNVs tested, however, this also limits the accuracy for each single region. The probe size ($<$100bp) is also much smaller than the tested region ($>$1000bp). Furthermore, due to the limitation of the probe design, the probes are not always placed in the middle of the region, so the issue may be due to CNV resolution. Second, in order to evaluate the contribution of SNP information, we repeated the analysis separately for CNV calls with SNPs and without SNPs. SNPs were found in all 11 deletion calls. In the 14 duplication calls, 10 had SNPs of which 9 were validated (90% validated); whereas 4 did not have SNPs of which 3 were validated (75% validated). The increased validation rate in duplications with SNPs suggests that ASReC improved detection accuracy. Finally, as a secondary analysis, we also compared AS-GENSENG calls in sample NA12272 with the NanoString estimated CNV calls in

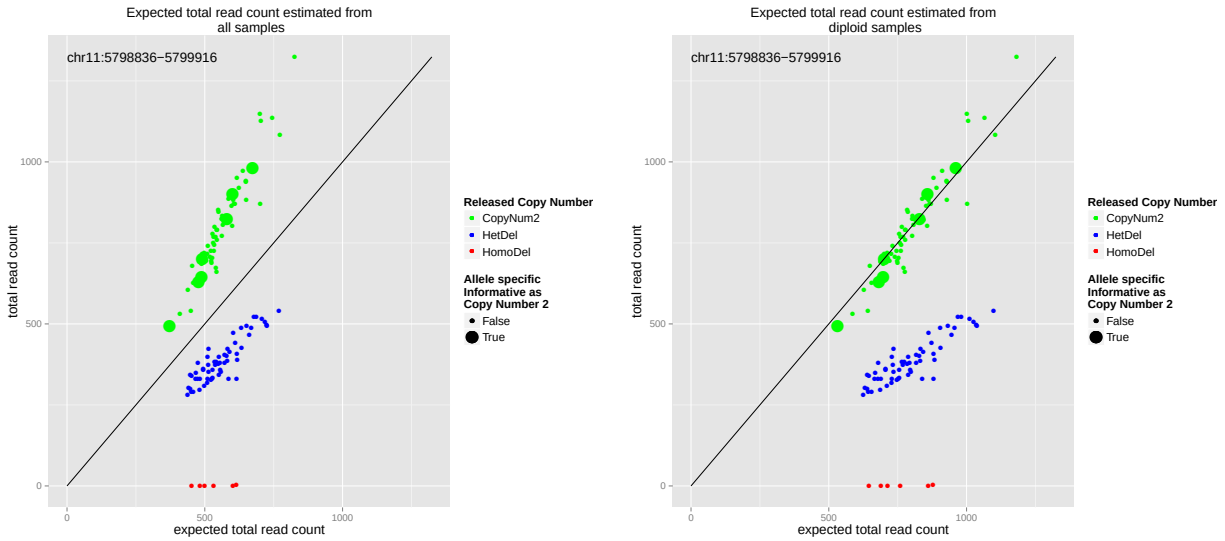
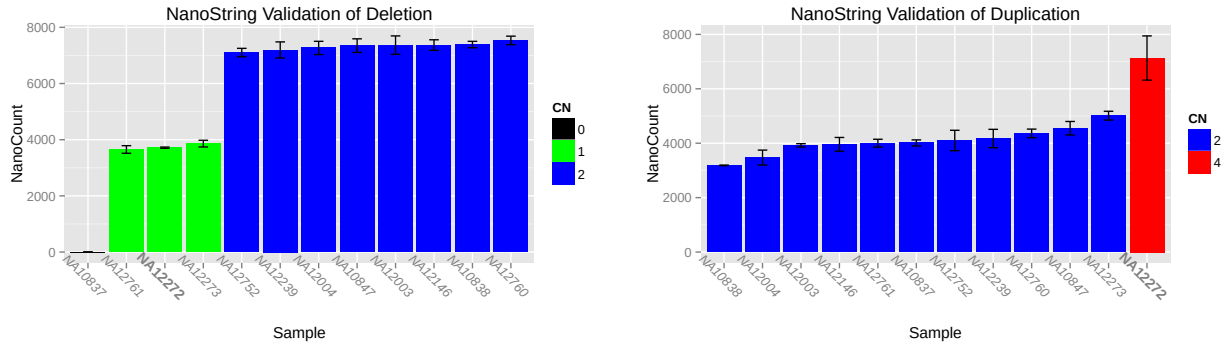


Figure 3.11: An example target that AS-GENSENG better estimates expected RC



(a) A validated deletion

(b) A validated duplication

Figure 3.12: Example: NanoString nCounter Technology Validation

the complete trio containing NA12272 (paternal: NA12272, maternal: NA12273, and child: NA10837). For each CNV, we looked for Mendelian inconsistencies in the NanoString copy number estimates. We found that among the 11 AS-GENSENG deletions, 10 were consistent. We found that all 12 AS-GENSENG duplications were consistent in this trio. If we assume that the NanoString copy number estimates in the parents were accurate, this analysis suggested 9% Mendelian error for deletions and 0% Mendelian error for duplications.

3.4 Discussion

We have developed an integrated and novel method (AS-GENSENG) that exploits the rich information of both read-depth (TReC) and allele-specific read-depth (ASReC) to detect CNVs and ASCNVs from both WGS data and WES data. We use HMM hidden states to model the underlying integer copy numbers, and combine the joint analysis of TReC and ASReC with simultaneous bias correction in data likelihood. The WGS module of AS-GENSENG can be applied to a single genome and its overall performance is better than several existing state-of-the-art methods. The WES module additionally uses ASReC to identify the copy number two reference group and leverages the large-scale nature of exome sequencing, critical for effective data normalization and accurate detection of common CNVs. To our knowledge, AS-GENSENG is the first tool capable of detecting ASCNVs from high-throughput sequencing data in germline DNA samples.

Analogous to the previous success with array-based CNV calling, we demonstrated that jointly using AS information with TReC allows not only the estimation of ASCN, but it also improves the estimation of total copy number (e.g., 1 copy deletion, 3 copy duplications). We show through numerous examples, using both WGS and WES data, that incorporating ASReC remarkably improves the performance of CNV detection. In addition, one novel component of AS-GENSENG is its beta-binomial distribution to model the allelic imbalance to incorporate allele-specific information. This approach, applied to model both deletions and duplications rather than only deletions, does not restrict the analysis to inbred genomes (Simpson et al., 2009) and does not require human effort to call ASCNV (Mayrhofer et al., 2013). We have also shown that using ASReC helps to accurately estimate the expected TReC of copy number two for a WES target, crucial for accurate CNV calling in WES data. Although previous studies (Krumm et al., 2012) have applied sophisticated analysis techniques to deal with the common CNV problem, we have shown that using ASReC is a novel and effective strategy to tackle this problem.

We aimed to conduct a comprehensive evaluation by comparing the performance of AS-GENSENG to multiple state-of-the-art methods. In order to provide an unbiased evaluation, we applied each method using its recommended parameters and quality control filters. Through

independent evaluations conducted by the 1000GP (Mills et al., 2011; Abecasis et al., 2012), Genome STRiP (Handsaker et al., 2011) was regarded as the best performing among existing methods for WGS data. Genome STRiP is a multi-sample method and requires at least 20 or 30 samples (`http://gatkforums.broadinstitute.org/discussion/1490/frequently-asked-questions`). In this study, we focused on detecting CNVs from a single genome and therefore did not compare AS-GENSENG with Genome STRiP. Finally, INDELs (insertions and deletions $<$ 50bps (Mills et al., 2011; Abecasis et al., 2012)) could not detected by AS-GENSENG and require specialized algorithms (Abecasis et al., 2012; Mills et al., 2006).

We used multiple approaches (i.e. simulation, SAV, trio-analysis, NanoString) to evaluate FDR as it is more challenging to estimate without the knowledge of true false negative CNVs in the genome. For AS-GENSENG, although the absolute FDR observed in real data is higher than that observed in simulation, the relative FDR is still lower than other methods under comparison. In this study, all analyses were performed in a high-throughput cluster-computing environment where each computing node had a shared memory of 48 GB. Sequencing data were split into individual chromosomes and chromosome-wise data were then analyzed in parallel on multiple computing nodes. Thus, the running time of a method is determined by the most time-consuming chromosome. Given read-depth data from WGS, AS-GENSENG can call CNVs for a sample with 30X coverage within 2 hours, while ERDS and CNVnator in $<$1 hour. For normalized read-depth data from WES, all three competing methods (AS-GENSENG, XHMM, Conifer) can call CNVs within 1 hour for 300 samples and 200K exon targets.

In summary, we have developed a novel method AS-GENSENG with the following distinguishing features: (a) Joint analysis of both TReC and ASReC while accounting for various experimental biases in sequence data. (b) Ability to detect both CNVs and ASCNVs from both WGS data and WES data. (c) Ability to leverage ASReC and large-scale nature of WES projects for effective data normalization and accurate detection of common CNVs with various frequencies. Through rigorous assessment using simulation, empirical data and independent technology, we

have demonstrated the superior performance of AS-GENSENG in numerous examples. We conclude that AS-GENSENG not only predicts accurate allele-specific CNV calls, but also improves the accuracy of total copy number calls.

CHAPTER 4: R-GENSENG

4.1 Introduction

As introduced in Section 1, GLM+NB methods do not achieve good scalability because the IRLS algorithm, the standard approach to fit GLMs, has a quadric complexity with the size of data. In addition, IRLS needs to be run multiple times until it converges. The randomized algorithm is a general computational strategy that has been widely studied by multiple disciplines, such as theoretic computer science and numerical linear algebra (Boyd, 2010). The basic idea is to randomly select a subset of data and solve the problem on the selected data with much reduced scale. The randomized algorithm is asymptotically faster than existing deterministic algorithms and is faster in numerical implementation in terms of clock time (Boyd, 2010; Halko et al., 2011). This feature is especially appealing in the problem of GLM+NB methods because of the quadric computational complexity of the IRLS algorithm (Drineas et al., 2006; Rokhlin and Tygert, 2008; Tygert, 2009; Avron et al., 2010; Drineas et al., 2010; Meng et al., 2014; Drineas et al., 2012; Ma et al., 2014). When sampling, the straightforward strategy is the uniform sampling strategy, which draws the sample uniformly at random. In contrast, the weighted sampling strategy (or probability sampling) draws the sample according to an importance sampling distribution. The choice between these two strategies varies in different applications (Ma et al., 2014). Here, we introduce the randomized GLM+NB coefficients estimator (RGE) for speeding up the GLM+NB based read-count analysis. Our RGE uses a weighted sampling strategy.

To illustrate the utility of RGE, we used the GLM+NB based CNV detection method GENSENG (Szatkiewicz et al., 2013) as an example and named the resulting RGE-GENSENG as "R-GENSENG". As described in Section 2, GENSENG implements a hidden Markov model (HMM) and the GLM+NB method to integrate bias correction and read-count analysis in a one-step

procedure. In GENSENG, the HMM emission probability describes the likelihood of the observed read-count data and is computed as a mixture of uniform distribution and the NB regression model (a form of GLM); therefore, multiple confounding factors (e.g. GC content and mappability) are simultaneously accounted for by including them as regression covariates and unknown sources of bias are accounted for by the NB dispersion parameter.

In this chapter, we first evaluated the consistency and the variance properties of RGE. We concluded that RGE is a consistent GLM+NB regression estimator, and that the weighting sampling strategy yields smaller regression coefficients estimation variance than using uniform sampling. We then performed simulation and real-data analysis to evaluate R-GENSENG in comparison to the original GENSENG. We concluded the R-GENSENG is ten times faster than the original GENSENG while maintaining GENSENG's accuracy in CNV detection. Taken together, our results suggest that RGE and the strategy developed in this work could be applied to other GLM+NB based read-count analyses in order to substantially improve their computational efficiency while preserving the analytic power.

4.2 Methods

In this section, we first introduce critical statistical properties of RGE concerning its consistency and variance and then introduce R-GENSENG - the application of RGE to GENSENG. We evaluated the consistency of RGE because RGE uses a subset of data points to estimate NB regression coefficients; and below we show that, as the number of data points used increases indefinitely, the resulting estimates converges in probability to the true values of the coefficients. We evaluated the variance of RGE because RGE applies a weighted sampling strategy to select the subset of data; and below we show that weighted sampling yield smaller estimation variance than uniform sampling.

4.2.1 The consistency of RGE

Following notations, we summarize the main theory in Theorem 4.1 .

We denote by $\mathbf{X} \in \mathbb{R}^{n \times p}$ the design matrix that is composed of n rows and p columns, and $\mathbf{y} \in \mathbb{R}^n$ the n-dimensional response vector. Let $\mathbf{x}_j = (x_{1j}, ..., x_{nj})^T$ be the j-th column of $\mathbf{X}$, and $x_{i,j} \in \mathbb{R}$ be the element at the i-th row and j-th column of $\mathbf{X}$. Let $\mathbf{X}^T$ be the transpose of $\mathbf{X}$.

Similar to the GLM settings defined in (Fan and Lv, 2011), we consider the response vector $\mathbf{y}$ that all its elements independently generated from an exponential family distribution with the density function

$$f_n(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}) \equiv \prod_{i=1}^{n} f_0(y_i; \theta_i, \phi) = \prod_{i=1}^{n} \left\{ c(y_i) \exp\left[\frac{y_i \theta_i - b(\theta_i)}{\phi} \right] \right\}$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ is an unknown p-dimsnional vector of regression coefficients, $\{f_0(y; \theta, \phi) : \theta \in \mathbb{R}\}$ is a family of distributions in the exponential family with dispersion parameter, $\phi \in (0, \infty)$, and $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)^T$ is the n-dimsnional vector of canonical parameter. Under the assumed GLM, the affine transformation of log-likelihood $\log f_n(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta})$ is given as

$$\ell_n(\boldsymbol{\beta}) = n^{-1} \left[\mathbf{y}^\mathbf{T} \boldsymbol{\theta} - \mathbf{1}^\mathbf{T} b(\boldsymbol{\theta}) \right].$$

Let the sampling indicator for the i-th entry, $i = 1, ..., n$ be

$$m_i = \begin{cases} 1 & \text{if } i\text{-th entry is sampled,} \\ 0 & \text{otherwise.} \end{cases}$$

The log likelihood of the sampled data $\tilde{\ell}_n(\boldsymbol{\beta})$ could be represented as

$$\tilde{\ell}_n(\boldsymbol{\beta}) = n^{-1} \left[(\mathbf{m} \circ \mathbf{y})^\mathbf{T} \boldsymbol{\theta} - \mathbf{m}^\mathbf{T} b(\boldsymbol{\theta}) \right].$$

The first derivative of $\tilde{\ell}_n(\boldsymbol{\beta})$ is

$$\tilde{\ell}'_n(\boldsymbol{\beta}) = n^{-1} \left[\mathbf{X}^\mathbf{T} (\mathbf{m} \circ \mathbf{y}) - \mathbf{X}^\mathbf{T} (\mathbf{m} \circ \mathbf{X}\boldsymbol{\beta}) \right] \tag{4.1}$$

where $\circ$ is the Hadamard (component wise) product.

Let $\boldsymbol{\beta}_0 = (\beta_{01}, ..., \beta_{0p})$ be the true coefficients. The maximum likelihood estimator can be estimated by setting the first derivative of $\tilde{\ell}'_n(\boldsymbol{\beta})$ (equation (4.1)) to be 0. We show that there exists a solution when equation (4.1) equals 0 that is inside the hypercube of the true coefficients.

Theorem 4.1. *For sufficient large n, there exists a solution $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ for equation (4.1) of $\mathbf{X^T}(\mathbf{m} \circ \mathbf{y}) - \mathbf{X^T}(\mathbf{m} \circ \mathbf{X}\boldsymbol{\beta}) = 0$ inside the hypercube*

$$\mathcal{N}_0 = \left\{ \boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta} - \boldsymbol{\beta}_0\|_\infty \leq d_n = O(n^{-\gamma_0}\sqrt{\log n}) \right\},$$

assuming $d_n \equiv 2^{-1} \min_{1 \leq j \leq p}\{|\beta_{0j}|\} = O(n^{-\gamma_0}\sqrt{\log n})$ for some $\gamma_0 \in (0, 1/2)$. Proof is as follows: Let

$$\begin{aligned}
\varphi(\boldsymbol{\delta}) &= -n\tilde{\ell}'_n(\boldsymbol{\beta}) = \mathbf{X^T}(\mathbf{m} \circ \mathbf{X}\boldsymbol{\delta}) - \mathbf{X^T}(\mathbf{m} \circ \mathbf{y}) \\
&= \mathbf{X^T}(\mathbf{m} \circ \mathbf{X}\boldsymbol{\delta}) - \mathbf{X^T}(\mathbf{m} \circ \mathbf{X}\boldsymbol{\beta_0}) - \left[\mathbf{X^T}(\mathbf{m} \circ \mathbf{y}) - \mathbf{X^T}(\mathbf{m} \circ \mathbf{X}\boldsymbol{\beta_0})\right] \\
&= \mathbf{X^T}(\mathbf{m} \circ \mathbf{X}\boldsymbol{\delta}) - \mathbf{X^T}(\mathbf{m} \circ \mathbf{X}\boldsymbol{\beta_0}) - \mathbf{X^T}[\mathbf{m} \circ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta_0})]
\end{aligned}$$

Thus prove Theorem 4.1 is equivalent to prove that there exists a solution inside the hypercube $\mathcal{N}$ to satisfy $\varphi(\boldsymbol{\delta}) = 0$.

Expanding $\mathbf{X^T}(\mathbf{m} \circ \mathbf{X}\boldsymbol{\delta})$ around $\boldsymbol{\beta}_0$ by the second order Taylor expansion we have,

$$\varphi(\boldsymbol{\delta}) = \left[\mathbf{X^T}\text{diag}(\mathbf{m})\mathbf{X}\right](\boldsymbol{\delta} - \boldsymbol{\beta}_0) - \boldsymbol{\xi} + \mathbf{r},$$

where $\boldsymbol{\xi} = (\xi_1, ..., \xi_p)^T = \mathbf{X^T}[\mathbf{m} \circ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta_0})] = \mathbf{X^T}\boldsymbol{\epsilon}$, $\mathbf{r} = (r_1, ..., r_p)^T$ are Lagrange reminders and for each $j = 1, ..., p$,

$$r_j = \frac{1}{2}(\boldsymbol{\delta} - \boldsymbol{\beta}_0)\nabla^2\left[\mathbf{x_j^T}(\mathbf{m} \circ \mathbf{X}\boldsymbol{\delta})\right](\boldsymbol{\delta} - \boldsymbol{\beta}_0),$$

It is straightforward to see the second derivative $\nabla^2 \left[\mathbf{x_j^T}(\mathbf{m} \circ \mathbf{X}\boldsymbol{\delta})\right] = 0$, so we have $\|\mathbf{r}\|_\infty = 0$.

Let

$$\bar{\varphi}(\boldsymbol{\delta}) = \left[\mathbf{X^T}\text{diag}\,(\mathbf{m})\,\mathbf{X}\right]^{-1} \varphi(\boldsymbol{\delta}) = \boldsymbol{\delta} - \boldsymbol{\beta}_0 + \mathbf{u},$$

where $\mathbf{u} = -\left[\mathbf{X^T}\text{diag}\,(\mathbf{m})\,\mathbf{X}\right]^{-1}[\boldsymbol{\xi} - \boldsymbol{r}]$. In order to obtain the range $\|\mathbf{u}\|_\infty$, we need to study the range of $\|\left[\mathbf{X^T}\text{diag}\,(\mathbf{m})\,\mathbf{X}\right]^{-1}\|_\infty$ and $\|\boldsymbol{\xi}\|_\infty$ (we already know that $\|\mathbf{r}\|_\infty = 0$), where the L_∞ norm of a matrix is the maximum of the L_1 norm of each row.

We study the ranges under our CNV problem setting using an illustration example which copy number and intercept are set as two covariates. Without the loss the generality we assume the j-th column ($j = 1, 2$) of the sampled data has been standardized such that $\overline{\mathbf{m} \circ \mathbf{x_j}} = 0$, and $\|\mathbf{m} \circ \mathbf{x_j}\|_2 = \sqrt{n_0}$, where n_0 is the size of the sampled data. If the copy number and the intercept have not been standardized, the conclusion still holds with $\|\mathbf{m} \circ \mathbf{x_j}\|_2$ assumed to be in the order of $\sqrt{n_0}$. $\left[\mathbf{X^T}\text{diag}\,(\mathbf{m})\,\mathbf{X}\right]$ becomes diag (n_0, n_0). The inverse is diag $\left(n_0^{-1}, n_0^{-1}\right)$ and the L_∞ norm is $n_0^{-1} = O(n^{-1})$.

Next we study $\|\boldsymbol{\xi}\|_\infty$ from probability perspective. We first define the event $\mathcal{E} = \left\{\|\boldsymbol{\xi}\|_\infty \leq c^{-1/2}\sqrt{n_0 \log n_0}\right\}$, where $\sqrt{n_0}$ is the L_2 norm of vector $\mathbf{m} \circ \mathbf{x_j}$. Equation (22) in (Fan and Lv, 2011) defines the following property,

$$P\left(|\mathbf{a^T Y} - \mathbf{a^T}\mu(\theta_\mathbf{0})| > \|\mathbf{a}\|_2 \varepsilon\right) \leq \psi\left(\varepsilon\right),$$

where $\mathbf{a}$ is a p dimension vector, $\varepsilon \in (0, \|\mathbf{a}\|_2/\|\mathbf{a}\|_\infty]$, $c = 1/(2v_0 + 2M)$ for some $M, v_0 \in (0, \infty)$, and $\psi\left(\varepsilon\right) = 2\exp^{-c\varepsilon^2}$. With this property, the probability that event $\mathcal{E}$ happens could be calculated as

$$\begin{aligned} P\left(\mathcal{E}\right) &\geq 1 - \sum_{j=1}^{p} P\left(|\boldsymbol{\xi}_j| \geq c^{-1/2}\sqrt{n_0 \log n_0}\right) \\ &\geq 1 - 2[pn_0^{-1}] \end{aligned}$$

The probability goes to 1 when n_0 goes to ∞. Thus the event $\mathcal{E}$ holds when n_0 goes to ∞. Thus $\|\boldsymbol{\xi}\|_\infty \leq c^{-1/2}\sqrt{n_0 \log n_0} = O(n^{1/2}\sqrt{\log n})$.

Based on the range of $\| \left[\mathbf{X^T}\text{diag}\,(\mathbf{m})\,\mathbf{X}\right]^{-1} \|_\infty$ and $\|\boldsymbol{\xi}\|_\infty$, we have

$$\begin{aligned}\|\mathbf{u}\|_\infty &\leq \| \left[\mathbf{X^T}\text{diag}\,(\mathbf{m})\,\mathbf{X}\right]^{-1} \|_\infty (\|\boldsymbol{\xi}\|_\infty + \|\mathbf{r}\|_\infty) \\ &= O(n^{-1/2}\sqrt{\log n}),\end{aligned}$$

so $\|\mathbf{u}\|_\infty = o(n^{-\gamma_0}\sqrt{\log n})$ for some $\gamma_0 \in (0, 1/2)$. Since for any $\boldsymbol{\delta} \in \mathcal{N}$, we have $\|\boldsymbol{\delta}\|_\infty \geq \|\boldsymbol{\beta}_0\|_\infty - d_n$, where $d_n \equiv 2^{-1}\min_{1\leq j\leq p}\{|\beta_{0j}|\} = O(n^{-\gamma_0}(\log n)^{1/2})$ for some $\gamma_0 \in (0, 1/2)$. Therefore we have

$$\min_{j=1,\ldots,p} \|\delta_j\| \geq \min_{j=1,\ldots,p} \|\boldsymbol{\beta}_{0j}\| - d_n = d_n.$$

For a constant $C > 0$ and sufficiently large n, if $\delta_j - \beta_j = Cn^{-\gamma_0}\sqrt{\log n}$, $\bar{\varphi}_j(\boldsymbol{\delta}) \geq Cn^{-\gamma_0}\sqrt{\log n} - \|\mathbf{u}\|_\infty \geq 0$. And if $\delta_j - \beta_j = -Cn^{-\gamma_0}\sqrt{\log n}$, $\bar{\varphi}_j(\boldsymbol{\delta}) \leq -Cn^{-\gamma_0}\sqrt{\log n} + \|\mathbf{u}\|_\infty \leq 0$. Because the continuity of function $\bar{\varphi}(\boldsymbol{\delta}) = (\bar{\varphi}_1(\boldsymbol{\delta}), ..., \bar{\varphi}_p(\boldsymbol{\delta}))$, and Miranda's existence theorem, there is a solution $\hat{\boldsymbol{\beta}}$ for $\varphi(\boldsymbol{\delta}) = 0$ in $\mathcal{N}$, i.e., there is a solution for Equation 4.1 in $\mathcal{N}$. Thus Theorem 4.1 holds.

4.2.2 The variance of RGE

RGE applies a weighted sampling strategy because of its potential in yielding estimation variance smaller than uniform sampling. Using a representative NB regression model, we evaluated and compared the inverses of the Fisher information matrix between RGEs weighted sampling and uniform sampling.

The co-variance matrix of the maximum likelihood estimator (MLE) $\boldsymbol{\beta}$ is the inverse of Fisher information matrix $-E(\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}^2})$. The Fisher information matrix is a $p \times p$ matrix, and its

(j,k)-th element equals to

$$-E\left(\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k}\right) = \sum_{i=1}^{n} \frac{1}{\text{Var}(y_i)} x_{ij} x_{ik},$$

if the link function is $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$.

The representative NB regression model used here is $\mu = \beta_0 + \beta_1(CN)$, where the link function is the identity link function, μ is the mean value of read-count, β_0 is the intercept, and β_1 is the coefficient of the copy number CN. CN measurements take three values: 0 for deletions, 1 for copy number neutral, and 2 for duplications. This model is representative because it covers the general characteristics of the read-count analysis: a biological factor (i.e. copy number) with three states including one state representing the baseline (i.e. copy number neutral) and two states representing the bidirectional differences from the baseline (i.e. deletions and duplications).

Under the studied regression model, the fisher information matrix is a 2×2 matrix considering the existence of intercept. The $(1,1)$ element is $\sum_{i=1}^{n} \frac{1}{\text{Var}(y_i)}$, the $(1,2)$ and the $(2,1)$ elements are $\sum_{i=1}^{n} \frac{1}{\text{Var}(y_i)} x_i$, and the $(2,2)$ element is $\sum_{i=1}^{n} \frac{1}{\text{Var}(y_i)} x_i^2$, where x_i is the copy number of the i-th observation. The inverse of a 2×2 matrix could be analytically given. Here we are interested in the variance of the coefficient of the copy number, which is the $(2,2)$ element of the inverse matrix. Define p_1 as the probability of deletion event happening, p_2 as the probability of copy number neutral happening, and p_3 as the probability of duplication happening. With the linear link function, the $(2,2)$ element equals

$$\frac{p_1 r + p_2 s + p_3 t}{n\left(p_1 p_2 r s + 4 p_1 p_3 r t + p_2 p_3 s t\right)}, \tag{4.2}$$

where $r = (\beta_0 + \phi\beta_0^2)^{-1}$, $s = (\beta_0 + \beta_1 + \phi(\beta_0 + \beta_1)^2)^{-1}$, and $t = (\beta_0 + 2\beta_1 + \phi(\beta_0 + 2\beta_1))^{-1}$.

From Eq. (4.2) we find that when the uniform sampling is applied, p_1, p_2 and p_3 would be the same in the sampled rows, but n would be smaller depending on the sampling size. As a result, the variance would become larger. For example, if we uniformly sample 10% of all rows, the

variance would be 10 times larger. Thus, the estimation of coefficients from the sampled data has larger variance.

We next compare the uniform sampling strategy with the weighted sampling strategy used in RGE by finding the minimum solution of Eq. (4.2) (i.e. the distribution of p_1, p_2 and p_3 in the sampled data that gave a minimum variance given the same sampling size). We list below the KKT-conditions for minimizing Eq. (4.2) subject to constraints. First, the objective function under the KKT-conditions is

$$\frac{p_1 r + p_2 s + p_3 t}{n\left(p_1 p_2 rs + 4 p_1 p_3 rt + p_2 p_3 st\right)} + \lambda(1 - p_1 - p_2 - p_3) - \mu_1 p_1 - \mu_2 p_2 - \mu_3 p_3,$$

where λ and μ_1, μ_2, and μ_3 are KKT multipliers. And the necessary conditions for the minimum solution are

Stationarity

$$\begin{aligned}
\frac{r(p_2 s + 2p_3 t)^2}{n(p_1 p_2 rs + 4p_1 p_2 rt + p_2 p_3 st)^2} &= \lambda + \mu_1, \\
\frac{s(p_1 r - p_3 t)^2}{n(p_1 p_2 rs + 4p_1 p_2 rt + p_2 p_3 st)^2} &= \lambda + \mu_2, \\
\frac{t(p_2 s + 2p_1 r)^2}{n(p_1 p_2 rs + 4p_1 p_2 rt + p_2 p_3 st)^2} &= \lambda + \mu_3.
\end{aligned}$$

Primal feasibility and Dual feasibility

$$p_1 + p_2 + p_3 = 1,$$

$$p_1 \geq 0, p_2 \geq 0, p_3 \geq 0,$$

$$\mu_1 \geq 0, \mu_2 \geq 0, \mu_3 \geq 0.$$

Complementary slackness

$$\mu_1 p_1 = 0, \mu_2 p_2 = 0, \mu_3 p_3 = 0.$$

Three possible solutions satisfy the KKT conditions.

$Solution1$

$p_1 = 0, p_2 = \frac{\sqrt{st}}{\sqrt{st}+s}, p_3 = \frac{\sqrt{s}}{\sqrt{s}+\sqrt{t}},$

objective function $= \frac{(\sqrt{1/s}+\sqrt{1/t})^2}{n}$

$Solution 2$

$p_1 = \frac{\sqrt{t}}{\sqrt{r}+\sqrt{t}}, p_2 = 0, p_3 = \frac{\sqrt{rt}}{\sqrt{rt}+t},$

objective function $= \frac{(\sqrt{1/r}+\sqrt{1/t})^2}{4n}$

$Solution 3$

$p_1 = \frac{\sqrt{s}}{\sqrt{r}+\sqrt{s}}, p_2 = \frac{\sqrt{rs}}{\sqrt{rs}+s}, p_3 = 0,$

objective function $= \frac{(\sqrt{1/r}+\sqrt{1/s})^2}{n}$

The objective function introduced above describes the scale of the inverse of the Fisher information matrix (i.e. the scale of estimation variance). We thus want to discuss when the minimal solution of the objective function could be achieved. Within the representative setting, $\mu = \beta_0 + \beta_1(CN)$, where CN is copy number from $0, 1, 2$. In this case, when $CN = 0$ (deletion), $\beta_0 = \mu$, where μ is the expected read count for deletion event. Thus $\beta_0 > 0$. On the other hand, the read count will increase with the copy number in a linear manner (i.e. the read count of copy number two region should be about twice of the read count of copy number one region), which suggests that the coefficient for CN β_1 should be close to 1. Given $\beta_0 > 0$ and $\beta_1 \simeq 1$, we have $1/r < 1/s < 1/t$, it is straightforward to see solution 3 is smaller than solution 1. We next compare solution 2 with solution 3. With a reasonable $\mu = 0.1$, we numerically solve the equation $\frac{(\sqrt{1/r}+\sqrt{1/t})^2}{4} < (\sqrt{1/r} + \sqrt{1/s})^2$ using the symbolic equation function in Matlab and conclude that solution 2 is the minimal solution. In solution 2, $p_2 = 0$, which means that if only sampling rows representing copy number variation the variance will be minimized.

So far, the variance studies above show two conclusions. The first is that when RGE selects a subset of data, the estimation variance will increase. The second is that if RGE only selects data representing copy number variations, the variance will be smaller. Empirically, data representing copy number variation is rare ($< 1\%$). Therefore, given a fixed sample size a weighted sampling strategy that assigns higher sampling probability to CNV data would select more data representing copy number variations (i.e. yielding a smaller estimation variance) than the uniform sampling. As a result, RGE decides to use a weighted sampling strategy.

4.2.3 RGE in a real application

In this section we demonstrate an example usage of RGE to speed up the analysis of read-count data in one real application. The application is detecting copy number variations (CNV) from the read-count data. GENSENG (Szatkiewicz et al., 2013) accurately detects CNV by distinguishing the true CNV signal and the false signals caused by confounding factors in the read-count data. Given a reference genome, GC-content is the proportion of G or C bases in each window; while mappability measures the uniqueness of sequence in each window (Szatkiewicz et al., 2013). These two experimental biases are the confounding factors, and GENSENG measures confounding factors effects by fitting a NB regression model with the read-count as the response and three covariates (the copy number, GC-content, and mappability). GENSENG implements a HMM to deal with that the copy number is a hidden variable. GENSENG applies an E-M procedure to iteratively estimate the most likely copy number for each window. In the E step, it calculates the emission probability from the regression coefficients estimated in the previous round; while in the M step, it runs IRLS to estimate NB regression coefficients. RGE reduces the running time by providing a much-reduced scale of data to IRLS in each M step. When sampling the data, RGE assigned a much higher sampling probability of 0.99 to copy number variation window (based on the current largest posterior probability of each window given the copy number); otherwise, RGE assigned 0.01. Below is the integration algorithm in the M step.

Algorithm 1 is applied in the M step before IRLS estimating the NB regression coefficients. The inputs of the algorithm include the response vector $\mathbf{y}$, the design matrix $\mathbf{X}$, which contains the read-count data of each window, and the values of three covariates at each window, respectively. The inputs also include IRLS weights (Green, 1984), which are the current posterior probabilities of the windows $p =< p_1, ..., p_n >$. Based on these weights, Algorithm 1 assigns sampling probability. The last but not the least input is the sampling size, parameter q. The output is the regression coefficients estimated on the subsampled data. After applying Algorithm 1 in the M step of GENSENG, we call the resulting method R-GENSENG.

Algorithm 1: Algorithm to integrate RGE with GENSNEG

Data: $\mathbf{X} \in \mathbb{R}^{n \times p}$,$\mathbf{y} \in \mathbb{R}^n$, $p =< p_1, ..., p_n >, p_i = (p_i 1, ..., p_i m)$,$q$.
Result: $\hat{\boldsymbol{\beta}}$
initialize a weights vector with length n all 0 $w =< w_1, ..., w_n >$;
for $i = 1$ **to** n **do**
    if *the largest item in* p_i *represents copy number neutral* **then**
        $w_i = 0.99$;
    else
        $w_i = 0.01$;
    $s = nq$;
    repeat
        generate random number $v \in (0, 1)$;
        sample idx row if $v < w_i$;
    until s *rows in* $\mathbf{X}$ *has been sampled*;
    denote sampled rows of the designed matrix as $\mathbf{X}' \in \mathbb{R}^{s \times p}$, sampled response vector as $\mathbf{y}' \in \mathbb{R}^s$;
    estimate $\hat{\boldsymbol{\beta}}$ using standard IRLS algorithm from GLM regressions with input $\mathbf{X}'$ and $\mathbf{y}'$;

4.3 Experiment Results

We conducted experiments to evaluate RGE and its application R-GENSENG. We evaluated the conclusions made in the consistency study and the variance study described in Section Methods. We also evaluated R-GENSENG's performance in detecting CNV from the read-count data and compared with GENSENG. The experiments were carried out on both simulation data and empirical data.

4.3.1 RGE validation

In the consistency study we claim that the coefficients estimated by RGE will asymptotically converge at their true values. The variance study claims that RGE yield smaller estimation variance than the uniform sampling. To evaluate these two properties, we simulated a read-count data set and applied RGE to it multiple times to obtain a population of coefficients estimations. We thus could study this distribution to determine RGE consistency and variance. To demonstrate its superiority, we further compared the RGE estimations with several different

estimation baselines (including ground truth coefficients). First lets introduce the read-count data simulation.

We simulated a population of read count data following the NB distribution, where the relationship between the response and three covariates is:

$$\mu \quad = \quad \beta_0 + \beta_1(CN) + \beta_2(l) + \beta_3(g) \tag{4.3}$$

where μ is the mean value of the read count data, CN is the copy number, l is the mappability score, g_t is the GC content. We first generated covariates data from human chromosome one (NCBI37). We split the chromosome into 200bps non-overlapping windows and then calculated the GC content and mappability (see Supplementary materials for details about GC content and mappability calculation). For covariate copy number, we empirically set 99% of windows to have copy number 2 (e.g., copy number neutral). For the other 1% of windows, we randomly assigned a copy number from 0,1 (representing deletions) or 36 (representing duplications). We selected 10^6 windows, which yields a $10^6 \times 3$ design matrix. We empirically set the values of coefficients $\beta_1, \beta_2, \beta_3$ as 1,1 and 0.55. After the covariates are generated, we passed the design matrix and the coefficients to the garsim function from R/gsarima to simulate read-count data that follow NB distribution. The settings details are as follows: Set the link function for the applied garsim model; set the zero correction parameter to "zq1", and set the inverse of the overdispersion parameter to 0.01.

After the read count data were generated, we applied RGE to it multiple times. Each run of RGE returns estimation on coefficients, and by aggregating estimations from multiple runs together we could determine the convergence and the variance of RGE by studying the their distribution. In addition, RGE allows specifying sampling proportion. We specified two sampling proportions: 10% and 50%, and we ran RGE 200 times with both sampling proportions. To demonstrate the superiority of RGE, we compared RGE estimations with the following three baselines: the ground truth coefficients $< 1, 1, 0.55 >$, the coefficients estimated using the entire dataset (10^6 rows), and the coefficients estimated using a uniformly sampled data.

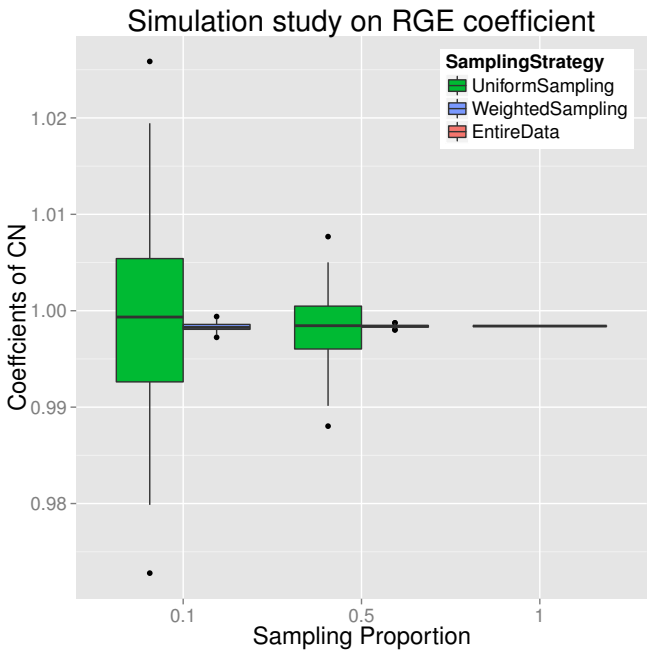


Figure 4.1: Simulation study of the coefficients estimated from sampled data

We plotted the estimations using a box plot, and show the estimations of CN in Figure 4.1. The x-axis represents the sampling proportion, and the y-axis represents the values of CN estimation. The RGE estimations are drawn as blue bars at x-axis values 0.1 and 0.5. Three baselines are represented by the y-axis with value 1 for the ground truth, a segment at x-axis value 1 for the estimation from the entire data, and two green bars at x-axis values 0.1 and 0.5 for uniform sampling, respectively.

From Figure 4.1 we have two observations. The first is that the RGE estimation converges at the ground truth. And the second is that RGE yields a smaller estimation variance than the uniform sampling. The experiment results strongly support that RGE is both consistent and with smaller coefficient estimation variance than the uniform sampling. Note that although the simulation experiments above were in CNV detection background; the conclusions are applicable in the more general GLM+NB based read-count analyses.

4.3.2 R-GENSENG performance evaluation

R-GENSENG is the usage of RGE to speed up the read-count based CNV detection analysis. We expected that R-GENSENG would be much faster than GENSENG. However, the coefficients estimated by RGE are pretty accurate. Therefore we expected the calls of R-GENSENG to be accurate and to be similar to the calls of GENSENG. We evaluated these two hypotheses with experiments carried out on both the simulation data and the empirical data. First we introduce the simulation evaluation.

Our simulation mimics the real scenario that detecting CNVs from sequencing data generated from a CNV-affected chromosome. The simulated sequencing reads were obtained from a hypothetical genome, in which we modified to implant CNVs. We first downloaded chromosome one, human reference genome (NCBI37) as a template, and then implanted 200 CNVs into it. A CNV is specified as starting position, ending position, and copy number. When the CNV type is duplication, we will copy the base pairs within the CNV region immediately after the CNV region. When the CNV type is deletion, we will remove the base pairs in the CNV region. After the CNV-implanted hypothetic genome was obtained, we applied the sequencing simulator, wgsim, as implemented in SAMTools (Li et al., 2009) to generate 100bps paired-end faked reads from it (specified as nucleic sequence in the form of a string, quality score, and so on). The reads were aligned back to unmodified human reference chromosome one (NCBI37).

Because window size affected read-count based CNV detection performance (Szatkiewicz et al., 2013). CNV will affect the read count of the windows in which it lies. When a CNV spans more windows, it will leave footage in more windows and thus will be more easily detected, especially for the HMM framework in which both GENSENG and R-GENSENG are applied. However, the sequencing data of noise and large window has the effect of washing off the noise. Thus we evaluated R-GENSENG performance systematically with read-count data generated with different window size. We evaluated with the window size ranging from 100bps, 200bps, and 500bps, up to 1000bps (see the Supplementary materials for counting method). We ran both GENSENG and R-GENSENG on these four sets of data. The set of implanted CNV is the ground

truth and we could calculate sensitivity and false discovery rate (FDR). Following (Szatkiewicz et al., 2013), a true discovery of sensitivity is an implanted CNV call that has more than 50% bases overlapped by predicted calls with the correct copy number type (both deletion or both duplication). A false discovery of FDR is a predicted call with less than 50% bases overlapped by implanted CNV calls. We compared the sensitivities and the FDRs of GENSENG and R-GENSENG. The results are summarized in tables 4.1 and 4.2.

To show the effect of CNV size on sensitivity, we stratified the sensitivity results by CNV size. Among the implanted CNVs there were 20 small CNVs (<1kbs), 86 median-size CNVs (between 1k and 3k bps), and 94 large CNVs (>3kbs). To summarize the sensitivity performance:

1) R-GENSENG detects almost the same number of implanted CNVs as GENSENG in the categories when a CNV spans several windows (e.g., a small-size CNV category when the window size is 100bps/200bps, or a median- or large-size CNV category for 500bps window size). When a CNV spans only a very small number of windows (e.g., a median-size CNV category when the window size is 1000bps), R-GENSENG detects a smaller number of implanted CNVs than GENSENG.

2) Smaller window size is extremely important for detecting small CNVs. When the window size is 100bps (a small CNV may span as many as 10 windows), 17/20 implanted small CNVs are detected by GENSENG and 15/20 CNVs are detected by R-GENSENG. But when the window size is 500bps (a small CNV could span only few windows), none of the implanted small CNVs is detected by either GENSENG or R-GENSENG.

A second aspect we examined is the FDR. Both GENSENG calls and R-GENSENG calls are accurate (low FDR) and GENSENG calls are slightly more accurate than R-GENSENG calls due to the formers use of all data. We also find that when the window size is small, there will be more false discovery (5.6% FDR of GENSENG when the window size is 100bps and 6.5% of R-GENSENG when the window size is 100bps) due to the difficulty of distinguishing a false signal caused by random noise from the true signal caused by CNV). While large window size has the

effect of washing off the noise, the FDRs are low when the window size is big ($< 5\%$ when the window size $>$ 100bps for GENSENG, and when the window size $>$ 100bps for R-GENSENG).

In short, based on the systematical evaluation results on simulation data, we could conclude that R-GENSENG is as accurate as GENSENG in terms of sensitivity and FDR.

Table 4.1: Sensitivity with different window size

Window Size	Implanted CNV by size (in bps)					
	$\leq$ 1k bps		$>$ 1k bps and $\leq$ 3k bps		$>$ 3k bps	
	20		86		94	
	G[1]	R[1]	G[1]	R[1]	G[1]	R[1]
100bps	17	15	81	81	90	90
200bps	16	12	81	81	90	90
500bps	0	0	80	79	89	89
1000bps	0	0	60	46	83	81

Table 4.2: FDR with different window size

Window Size	FDR	
	GENSENG	*R-GENSENG*
100bps	11/199(5.6%)	13/199(6.5%)
200bps	7/194(3.6%)	9/192 (4.7%)
500bps	5/174(2.9%)	7/175(4%)
1000bps	4/147(2.7%)	6/133(4.5%)

We then evaluated R-GENSENG empirically on the whole-genome sequencing data from 3 HapMap individuals sequenced as part of the 1000 Genomes Project (Abecasis et al., 2012; Mills et al., 2011). These data included the CEU parent-offspring trio of European ancestry (NA12878, NA12891, NA12892), sequenced to 40X coverage on average using the Illumina Genome Analyzer (I and II) platform. Sequencing reads are a mixture of single-end and paired-end with variable lengths (36bp, 51bp). The complete genome sequence data were obtained in the form of .bam alignment files from `ftp://ftp.ncbi.nlm.nih.gov/1000genomes/ftp/pilot_data/data/`. Similar

[1]G: GENSENG; R: R-GENSENG

to the procedure in (Szatkiewicz et al., 2013), we preprocessed the data by QC (removing reads with q-value < 10).

We also systematically evaluated with different window size. We also applied 100bps, 200bps, 500bps, and 1000bps non-overlapping windows. Thus we had 12 total sets of read count data, 4 for each sample. We then applied both GENSENG and R-GENSENG (sampling 10% of data) to these data.

Different window size brings different data size. To evaluate the efficiency of R-GENSENG, we recorded the running time on different time in seconds. For a particular window size of a sample, we summed up the running time on entire autosomes. We then for each window size calculated the average running time over the three samples. We present the running time result in Figure 4.2. The x-axis is the window size and the y-axis is the running time in seconds. There are two curves in the figure. The red curve connects the running time of GENSENG with different window sizes and the blue curve connects the running time of R-GENSENG. From Figure 4.2 we have two findings. The first is that the red curve is always high above the blue curve, which indicates that R-GENSENG runs much faster than GENSENG. In fact, R-GENSENG is nearly one magnitude faster than GENSENG. The second is that when the window size is small (100bps), the data size is huge and the absolute advantage of R-GENSENG over GENSENG is remarkable (R-GENSENG uses 6 hours vs. GENSENG uses 60 hours).

The second aspect we examined is the accuracy of R-GENSENG calls. We previously applied GENSENG to the trio data and found the calls to be accurate with high sensitivity and low FDR (Szatkiewicz et al., 2013). With GENSENG calls as a baseline, we intersected the R-GENSENG calls with the GENSENG calls using a 50% reciprocal overlap condition and reported the overlapping percentage. We expected that R-GENSENG would have a high overlapping percentage with GENSENG calls. We present the overlapping percentage of R-GENSENG calls in Figure 4.3. Each row represents the intersection results for three samples of a particular window and each column is stratified by CNV type, including overall (CNV), deletion,

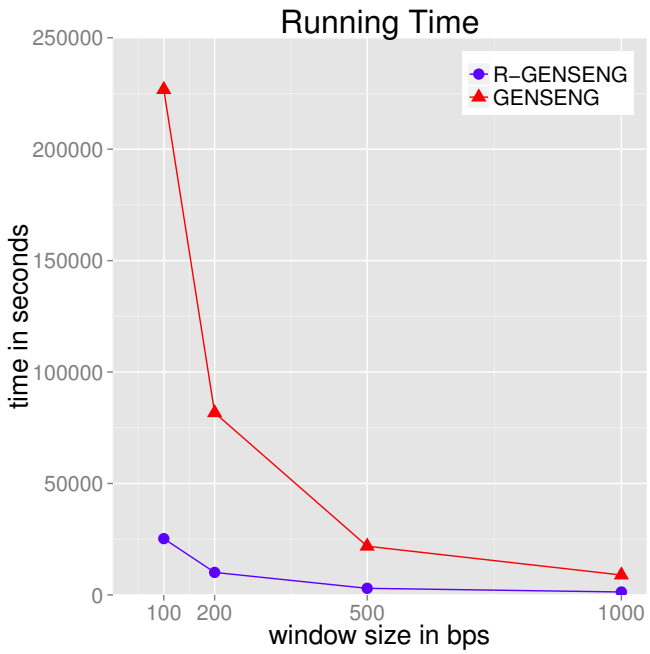


Figure 4.2: Running time of the real data with different window sizes

and duplication. We use the height of the bar to present the overlapping percentage, and color to denote sample. We find that the overlapping percentages are > 0.9 for most of the cases.

In short, R-GENSENG not only runs much faster than GENSENG, but also produces calls as accurately as GENSENG does. It proves the success of applying RGE in the CNV detection application.

4.4 Discussion

The GLM+NB methods are widely applied in the analysis of genomic read-count data produced by HTS technologies. However, the computation burden of the GLM+NB methods hinders their applications to the analysis of large-scale HTS data generated everyday. In this chapter, we have proposed a randomized GLM+NB estimator RGE that efficiently estimates the regression coefficients. We showed the accuracy of RGE by proving that it is consistent and has smaller coefficient estimation variance. We also conducted simulation experiments to evaluate RGEs consistency and variance properties. RGE has not been seen in the existing randomized algorithms literatures (Drineas et al., 2006; Rokhlin and Tygert, 2008; Tygert, 2009; Drineas et al.,

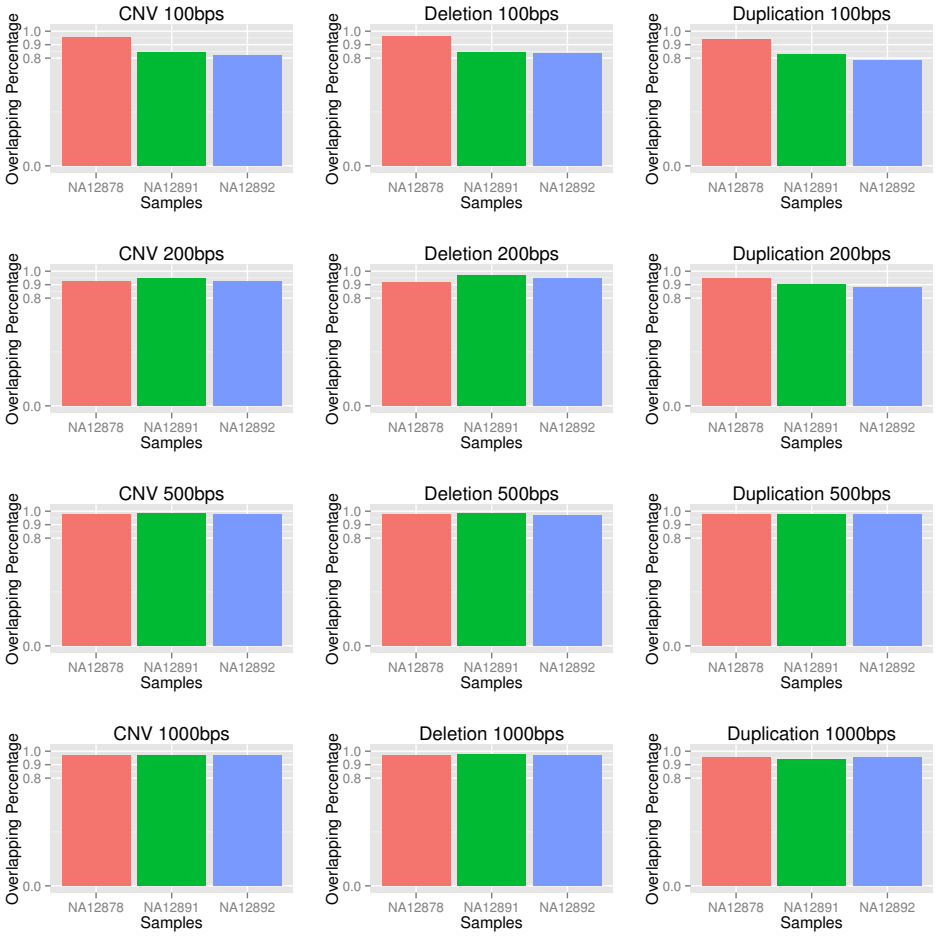


Figure 4.3: Overlapping percentage of R-GENSENG calls in empirical data

2010), and is applicable for speeding up the general read-count data analysis that applies GLM+NB methods. We showed in this chapter one application of RGE in speeding up a read-count based CNV detection framework GENSENG (Szatkiewicz et al., 2013). The resulting method is named as R-GENSENG. We demonstrated the success of R-GENSENG using experiments carried out on both simulation data and empirical data. We found that R-GENSENG not only runs much faster than GENSENG but also produces accurate CNV calls.

CHAPTER 5: CONCLUSIONS

The dissertation presents an effective yet efficient integrated likelihood-based CNV detection framework for HTS data. With the sequentially placed read-depth data, the framework employs a HMM to exploit the spatial information, i.e. the consecutive windows tend to have the same copy number. The framework applies a negative-binomial regression (a special case of GLM) to estimate the relation among expected read count with the copy number and covariates from the read-depth data. The negative-binomial distribution allows the variance be larger than mean, which better fits the real data. On the other hand, the framework further incorporates the allele-specific information provided by allele-specific reads. With the intention to study simple CNV (no two CNV happens at the same place), there are different allelic configurations defined for each copy number, and the framework models the likelihood of each allelic configuration using a binomial distribution while the likelihood of each copy number is the weighted average of the associated allelic configurations. The emission probability is generated based on the product of the likelihood of read-depth and the likelihood of allele-specific information. HMM produces the posterior probability at each window for each defined copy number. The proposed framework segments CNV based on the posterior probability. The framework also generates ASCN after the CNV segmentation by collecting allele-specific information inside each CNV and predict the ASCN as the most likely allelic configuration of the called CNV. In order to improve the efficiency of the framework, this dissertation proposes a randomized GLM estimator RGE to tackle the computation bottleneck of the framework: GLM+NB methods. RGE follows the general idea of sampling to reduce the scale of data to speed up the analysis. However, it is novel because no randomized algorithm has been conducted on the LGM+NB method. The coefficient estimated by RGE is consistent and yields smaller coefficient estimation variance. We also show its application in GENSENG, and conduct experiments to evaluate the performance of R-GENSENG.

This dissertation presents extensive experiments conducted to validate the performance of the proposed CNV detection framework, and throughly comparison with the state of the art peer CNV detection methods (CNVnator (Abyzov et al., 2011), ERDS (Heinzen et al., 2012), XHMM (Fromer et al., 2012), Conifer (Krumm et al., 2012), etc.). The experiment data presented in this dissertation come from WGS data to WES data, from human data to mouse data, from real data to simulation data, and from well calibrated data to exploring data. Based on the comprehensive results, it is concluded here that the proposed framework has achieved high accuracy and efficiency in CNV detection from the read-depth of HTS data. In this study, all analyses were performed in a high-throughput cluster-computing environment where each computing node had a shared memory of 48 GB. Sequencing data were split into individual chromosomes and chromosome-wise data were then analyzed in parallel on multiple computing nodes. Thus, the running time of a method is determined by the most time-consuming chromosome. Given read-depth data from WGS, AS-GENSENG can call CNVs for a sample with $\sim 30X$ coverage within 2 hours, while ERDS and CNVnator in < 1 hour. For normalized read-depth data from WES, all three competing methods (AS-GENSENG, XHMM, Conifer) can call CNVs within 1 hour for 300 samples and 200K exon targets.

With the success of the proposed framework in this dissertation, there are several future directions that needs researcher's further exploration.

Not all the CNVs can be detected by read-depth data. Upon examining the high-confidence CNVs from 1000GP (Mills et al., 2011), we found that $\sim$76% of high-confidence deletions and only $\sim$21% of high-confidence duplications were read-depth accessible from the 1000GP HTS data using 36bp and 51bp reads. The percentage of read-depth accessible regions may increase for longer reads and when we incorporate reads that are mapped to multiple locations in the genome. On the other hand, the high confidence CNVs may be inaccurate. For example, undetected CNVs in a reference individual can lead to mistaken copy number calls in the study samples (Abyzov et al., 2011; Yoon et al., 2009; Park et al., 2010).

Duplications are generally more challenging to detect than deletions by read-depth-based methods for several reasons (Alkan et al., 2011; Abyzov et al., 2011; Baker, 2012). First, the read-depth distribution (Poisson or negative binomial) suggests that the higher the read-depth signal the larger the signal variance. As expected, read-depth-based methods suffer reduced sensitivity in the detection of duplications (higher variance) compared with deletions (lower variance). Second, as mentioned in the proceeding paragraph, the proportion of read-depth-accessible high-confidence duplications is much less than that for deletions (21% vs 76%), thus reducing the sensitivity (Table 2.4). Lastly, as noted by (Abyzov et al., 2011), abnormally high read-depth signal may not necessarily represent a true duplication but rather the effect of an unknown reference. Quality control procedures that are aimed to reduce such false positive duplications (e.g. removing windows that are read-depth outlier or have any overlap with known genomic gaps) would lead to reduced sensitivity for duplications overall.

We are aware of several limitations with the read-depth approach and recommend alternative strategies. First, we focused on the accurate detection of simple CNVs and computed TReC using reads with unambiguous mapping in the reference genome. This approach results in lower power to detect complex CNVs within repeated sequences. For detecting CNVs in repeat-rich region, we recommend the use of specialized methods that are capable of considering all mapping positions and handling the uncertainty of read mapping (Sudmant et al., 2010; Wang et al., 2013; He et al., 2011; Alkan et al., 2009; Hormozdiari et al., 2009, 2010). Second, the WGS module of our method used a sliding window approach to compute TReC and ASReC, which results in lower power in detecting CNVs that are smaller than 1Kb. For deletions $<$1Kb, we recommend ERDS (Heinzen et al., 2012) or Genome STRiP (Handsaker et al., 2011) as these methods further utilize read-pair information for improved detection. Third, while our WES module is robust against CNV frequency, its power for detecting rare exonic CNVs is lower than methods that are optimized for this class of variants. In this paradigm, XHMM appears to have superior sensitivity for detecting rare CNVs from WES data and the quality score provided by XHMM could be informative in downstream analyses in order to improve specificity (Fromer et al., 2012).

We aimed to conduct a comprehensive evaluation by comparing the performance of AS-GENSENG to multiple state-of-the-art methods. In order to provide an unbiased evaluation, we applied each method using its recommended parameters and quality control filters. Through independent evaluations conducted by the 1000GP, Genome STRiP (Handsaker et al., 2011) was regarded as the best performing among existing methods for WGS data. Genome STRiP is a multi-sample method and requires at least 20 or 30 samples (`http://gatkforums.broadinstitute.org/discussion/1490/frequently-asked-questions`). In this study, we focused on detecting CNVs from a single genome and therefore did not compare AS-GENSENG with Genome STRiP. We used multiple approaches (i.e. simulation, SAV, trio-analysis, NanoString) to evaluate FDR as it is more challenging to estimate without the knowledge of true false negative CNVs in the genome. For AS-GENSENG, although the absolute FDR observed in real data is higher than that observed in simulation, the relative FDR is still lower than other methods under comparison.

REFERENCES

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. a., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., and McVean, G. a. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome research*, 21(6):974–84.

Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature reviews. Genetics*, 12(5):363–76.

Alkan, C., Kidd, J. M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., Kitzman, J. O., Baker, C., Malig, M., Mutlu, O., Sahinalp, S. C., Gibbs, R. a., and Eichler, E. E. (2009). Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature genetics*, 41(10):1061–7.

Amarasinghe, K. C., Li, J., and Halgamuge, S. K. (2013). CoNVEX: copy number variation estimation in exome sequencing data using HMM. *BMC bioinformatics*, 14 Suppl 2(Suppl 2):S2.

Amarasinghe, K. C., Li, J., Hunter, S. M., Ryland, G. L., Cowin, P. a., Campbell, I. G., and Halgamuge, S. K. (2014). Inferring copy number and genotype in tumour exome data. *BMC genomics*, 15:732.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, 11:R106.

Attiyeh, E. F., Diskin, S. J., Attiyeh, M. a., Mossé, Y. P., Hou, C., Jackson, E. M., Kim, C., Glessner, J., Hakonarson, H., Biegel, J. a., and Maris, J. M. (2009). Genomic copy number determination in cancer cells from single nucleotide polymorphism microarrays based on quantitative genotyping corrected for aneuploidy. *Genome research*, 19(2):276–83.

Avron, H., Maymounkov, P., and Toledo, S. (2010). Blendenpik: Supercharging LAPACK's Least-Squares Solver.

Baker, M. (2012). Structural variation: the genome's hidden architecture.

Balikova, I., Lehesjoki, A. E., De Ravel, T. J. L., Thienpont, B., Chandler, K. E., Clayton-Smith, J., Träskelin, A. L., Fryns, J. P., and Vermeesch, J. R. (2009). Deletions in the VPS13B (COH1) gene as a cause of Cohen syndrome. *Human Mutation*, 30(9).

Baum, L., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics*.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M. J., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara E Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G.-D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R., and Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59.

Bilmes, J. A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models.

Bochukova, E. G., Huang, N., Keogh, J., Henning, E., Purmann, C., Blaszczyk, K., Saeed, S., Hamilton-Shield, J., Clayton-Smith, J., O'Rahilly, S., Hurles, M. E., and Farooqi, I. S. (2010). Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature*, 463(7281):666–670.

Boyd, M. W. M. (2010). Randomized Algorithms for Matrices and Data. *Foundations and Trends in Machine Learning*, 3(2):123–224.

Brahmachary, M., Guilmatre, A., Quilez, J., Hasson, D., Borel, C., Warburton, P., and Sharp, A. J. (2014). Digital genotyping of macrosatellites and multicopy genes reveals novel biological

functions associated with copy number variation of large tandem repeats. *PLoS genetics*, 10(6):e1004418.

Cahan, P., Li, Y., Izumi, M., and Graubert, T. A. (2009). The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells. *Nature genetics*, 41(4):430–437.

Campbell, P. J., Stephens, P. J., Pleasance, E. D., O'Meara, S., Li, H., Santarius, T., Stebbings, L. A., Leroy, C., Edkins, S., Hardy, C., Teague, J. W., Menzies, A., Goodhead, I., Turner, D. J., Clee, C. M., Quail, M. A., Cox, A., Brown, C., Durbin, R., Hurles, M. E., Edwards, P. A. W., Bignell, G. R., Stratton, M. R., and Futreal, P. A. (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature genetics*, 40(6):722–729.

Chen, M., Gunel, M., and Zhao, H. (2013). SomatiCA: identifying, characterizing and quantifying somatic copy number aberrations from cancer genome sequencing data. *PloS one*, 8(11):e78143.

Chiang, D. Y., Getz, G., Jaffe, D. B., O'Kelly, M. J. T., Zhao, X., Carter, S. L., Russ, C., Nusbaum, C., Meyerson, M., and Lander, E. S. (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature methods*, 6(1):99–103.

Clop, a., Vidal, O., and Amills, M. (2012). Copy number variation in the genomes of domestic animals. *Animal genetics*, 43(5):503–17.

Coin, L. J. M., Cao, D., Ren, J., Zuo, X., Sun, L., Yang, S., Zhang, X., Cui, Y., Li, Y., Jin, X., and Wang, J. (2012). An exome sequencing pipeline for identifying and genotyping common CNVs associated with disease with application to psoriasis. *Bioinformatics (Oxford, England)*, 28(18):i370–i374.

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T. D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C. H., Kristiansson, K., Macarthur, D. G., Macdonald, J. R., Onyiah, I., Pang, A. W. C., Robson, S., Stirrups, K., Valsesia, A., Walter, K., Wei, J., Tyler-Smith, C., Carter, N. P., Lee, C., Scherer, S. W., and Hurles, M. E. (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–712.

Cooper, G. M., Coe, B. P., Girirajan, S., Rosenfeld, J. A., Vu, T. H., Baker, C., Williams, C., Stalker, H., Hamid, R., Hannig, V., Abdel-Hamid, H., Bader, P., McCracken, E., Niyazov, D., Leppig, K., Thiese, H., Hummel, M., Alexander, N., Gorski, J., Kussmann, J., Shashi, V., Johnson, K., Rehder, C., Ballif, B. C., Shaffer, L. G., and Eichler, E. E. (2011). A copy number variation morbidity map of developmental delay. *Nature genetics*, 43(9):838–846.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5):491–498.

Doco, T., Wieruszeski, J. M., Fournet, B., Carcano, D., Ramos, P., and Loones, A. (1990). Structure of an exocellular polysaccharide produced by Streptococcus thermophilus. *Carbohydrate research*, 198(2):313–321.

Drineas, P., Magdon-Ismail, M., Mahoney, M. W., and Woodruff, D. P. (2012). Fast approximation of matrix coherence and statistical leverage. *The Journal of Machine Learning Research*, 13(1):3475–3506.

Drineas, P., Mahoney, M. W., and Muthukrishnan, S. (2006). Sampling algorithms for l 2 regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm - SODA '06*, pages 1127–1136, New York, New York, USA. ACM Press.

Drineas, P., Mahoney, M. W., Muthukrishnan, S., and Sarlós, T. (2010). Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality.

Fanciulli, M., Norsworthy, P. J., Petretto, E., Dong, R., Harper, L., Kamesh, L., Heward, J. M., Gough, S. C. L., de Smith, A., Blakemore, A. I. F., Froguel, P., Owen, C. J., Pearce, S. H. S., Teixeira, L., Guillevin, L., Graham, D. S. C., Pusey, C. D., Cook, H. T., Vyse, T. J., and Aitman, T. J. (2007). FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nature genetics*, 39(6):721–723.

Fromer, M., Moran, J. L., Chambert, K., Banks, E., Bergen, S. E., Ruderfer, D. M., Handsaker, R. E., McCarroll, S. a., O'Donovan, M. C., Owen, M. J., Kirov, G., Sullivan, P. F., Hultman, C. M., Sklar, P., and Purcell, S. M. (2012). Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *American journal of human genetics*, 91(4):597–607.

Gamazon, E. R., Cox, N. J., and Davis, L. K. (2014). Structural architecture of SNP effects on complex traits. *American journal of human genetics*, 95(5):477–89.

Gardina, P. J., Lo, K. C., Lee, W., Cowell, J. K., and Turpaz, Y. (2008). Ploidy status and copy number aberrations in primary glioblastomas defined by integrated analysis of allelic ratios, signal ratios and loss of heterozygosity using 500K SNP Mapping Arrays. *BMC genomics*, 9:489.

Geiss, G. K., Bumgarner, R. E., Birditt, B., Dahl, T., Dowidar, N., Dunaway, D. L., Fell, H. P., Ferree, S., George, R. D., Grogan, T., James, J. J., Maysuria, M., Mitton, J. D., Oliveri, P., Osborn, J. L., Peng, T., Ratcliffe, A. L., Webster, P. J., Davidson, E. H., Hood, L., and Dimitrov, K. (2008). Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nature biotechnology*, 26(3):317–325.

Green, P. J. (1984). Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(2):149–192.

Greenman, C. D., Bignell, G., Butler, A., Edkins, S., Hinton, J., Beare, D., Swamy, S., Santarius, T., Chen, L., Widaa, S., Futreal, P. A., and Stratton, M. R. (2010). PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics (Oxford, England)*, 11(1):164–75.

Guo, Y., Sheng, Q., Samuels, D. C., Lehmann, B., Bauer, J. A., Pietenpol, J., and Shyr, Y. (2013). Comparative study of exome copy number variation estimation tools using array comparative genomic hybridization as control. *BioMed research international*, 2013:915636.

Guryev, V., Saar, K., Adamovic, T., Verheul, M., van Heesch, S. A. A. C., Cook, S., Pravenec, M., Aitman, T., Jacob, H., Shull, J. D., Hubner, N., and Cuppen, E. (2008). Distribution and functional impact of DNA copy number variation in the rat. *Nature genetics*, 40(5):538–545.

Halko, N., Martinsson, P. G., and Tropp, J. A. (2011). Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Review*, 53(2):217–288.

Handsaker, R. E., Korn, J. M., Nemesh, J., and McCarroll, S. a. (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nature genetics*, 43(3):269–76.

He, D., Hormozdiari, F., Furlotte, N., and Eskin, E. (2011). Efficient algorithms for tandem copy number variation reconstruction in repeat-rich regions. *Bioinformatics (Oxford, England)*, 27(11):1513–20.

Heinzen, E., Feng, S., Maia, J., He, M., Ruzzo, E., Need, A., Shianna, K., Pelak, K., Han, Y., Goldstein, D., Gumbs, C., Singh, A., Zhu, Q., Ge, D., Cirulli, E., and Zhu, M. (2012). Using ERDS to Infer Copy-Number Variants in High-Coverage Genomes.

Henrichsen, C. N., Vinckenbosch, N., Zöllner, S., Chaignat, E., Pradervand, S., Schütz, F., Ruedi, M., Kaessmann, H., and Reymond, A. (2009). Segmental copy number variation shapes tissue transcriptomes. *Nature genetics*, 41(4):424–429.

Holt, C., Losic, B., Pai, D., Zhao, Z., Trinh, Q., Syam, S., Arshadi, N., Jang, G. H., Ali, J., Beck, T., McPherson, J., and Muthuswamy, L. B. (2014). WaveCNV: Allele-specific copy number alterations in primary tumors and xenograft models from next-generation sequencing. *Bioinformatics*, 30(6):768–774.

Hormozdiari, F., Alkan, C., Eichler, E. E., and Sahinalp, S. C. (2009). Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Research*, 19(7):1270–1278.

Hormozdiari, F., Hajirasouliha, I., Dao, P., Hach, F., Yorukoglu, D., Alkan, C., Eichler, E. E., and Sahinalp, S. C. (2010). Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics (Oxford, England)*, 26(12):i350–7.

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W., and Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature genetics*, 36(9):949–951.

Iskow, R. C., Gokcumen, O., Abyzov, A., Malukiewicz, J., Zhu, Q., Sukumar, A. T., Pai, A. A., Mills, R. E., Habegger, L., Cusanovich, D. A., Rubel, M. A., Perry, G. H., Gerstein, M., Stone, A. C., Gilad, Y., and Lee, C. (2012). Regulatory element copy number differences shape primate expression profiles.

Ivakhno, S., Royce, T., Cox, A. J., Evers, D. J., Cheetham, R. K., and Tavaré, S. (2010). CNAseg-a novel framework for identification of copy number changes in cancer from second-generation sequencing data. *Bioinformatics*, 26(24):3051–3058.

Juang, B.-H. and Rabiner, L. R. (1985). Mixture autoregressive hidden Markov models for speech signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(6):1404–1413.

Karakoc, E., Alkan, C., O'Roak, B. J., Dennis, M. Y., Vives, L., Mark, K., Rieder, M. J., Nickerson, D. A., and Eichler, E. E. (2012). Detection of structural variants and indels within exome data. *Nature methods*, 9(2):176–8.

Keane, T. M., Goodstadt, L., Danecek, P., White, M. A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., Furlotte, N. A., Eskin, E., Nellå ker, C., Whitley, H., Cleak, J., Janowitz, D., Hernandez-Pliego, P., Edwards, A., Belgard, T. G., Oliver, P. L., McIntyre, R. E., Bhomra, A., Nicod, J., Gan, X., Yuan, W., van der Weyden, L., Steward, C. A., Bala, S., Stalker, J., Mott, R., Durbin, R., Jackson, I. J., Czechanski, A., Guerra-Assunção, J. A., Donahue, L. R., Reinholdt, L. G., Payseur, B. A., Ponting, C. P., Birney, E., Flint, J., and Adams, D. J. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, 477(7364):289–294.

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, N. A., Tsang, P., Newman, T. L., Tüzün, E., Cheng, Z., Ebling, H. M., Tusneem, N., David, R., Gillett, W., Phelps, K. A., Weaver, M., Saranga, D., Brand, A., Tao, W., Gustafson, E., McKernan, K., Chen, L., Malig, M., Smith, J. D., Korn, J. M., McCarroll, S. A., Altshuler, D. A., Peiffer, D. A., Dorschner, M., Stamatoyannopoulos, J., Schwartz, D., Nickerson, D. A., Mullikin, J. C., Wilson, R. K., Bruhn, L., Olson, M. V., Kaul, R., Smith, D. R., and Eichler, E. E. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191):56–64.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., and Wilson, R. K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, 22(3):568–76.

Korn, J. M., Kuruvilla, F. G., McCarroll, S. A., Wysoker, A., Nemesh, J., Cawley, S., Hubbell, E., Veitch, J., Collins, P. J., Darvishi, K., Lee, C., Nizzari, M. M., Gabriel, S. B., Purcell, S., Daly, M. J., and Altshuler, D. (2008). Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature genetics*, 40(10):1253–1260.

Krumm, N., Sudmant, P. H., Ko, A., O'Roak, B. J., Malig, M., Coe, B. P., Quinlan, A. R., Nickerson, D. a., and Eichler, E. E. (2012). Copy number variation detection and genotyping from exome sequence data. *Genome research*, 22(8):1525–32.

Laird, P. W. (2010). Principles and challenges of genomewide DNA methylation analysis. *Nature reviews. Genetics*, 11(3):191–203.

Levinson, D. F., Duan, J., Oh, S., Wang, K., Sanders, A. R., Shi, J., Zhang, N., Mowry, B. J., Olincy, A., Amin, F., Cloninger, C. R., Silverman, J. M., Buccola, N. G., Byerley, W. F., Black, D. W., Kendler, K. S., Freedman, R., Dudbridge, F., Pe'er, I., Hakonarson, H., Bergen, S. E., Fanous, A. H., Holmans, P. A., and Gejman, P. V. (2011). Copy number variants in schizophrenia: confirmation of five previous findings and new evidence for 3q29 microdeletions and VIPR2 duplications. *The American journal of psychiatry*, 168(3):302–316.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–60.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–9.

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11):1851–1858.

Li, J., Lupat, R., Amarasinghe, K. C., Thompson, E. R., Doyle, M. A., Ryland, G. L., Tothill, R. W., Halgamuge, S. K., Campbell, I. G., and Gorringe, K. L. (2012). CONTRA: copy number analysis for targeted resequencing. *Bioinformatics (Oxford, England)*, 28(10):1307–13.

Liu, E. Y., Li, M., Wang, W., and Li, Y. (2013). MaCH-admix: genotype imputation for admixed populations. *Genetic epidemiology*, 37(1):25–37.

Love, M. I., Myšičková, A., Sun, R., Kalscheuer, V., Vingron, M., and Haas, S. A. (2011). Modeling read counts for CNV detection in exome sequencing data. *Statistical applications in genetics and molecular biology*, 10(1).

Ma, P., Mahoney, M., and Yu, B. (2014). A Statistical Perspective on Algorithmic Leveraging. *JMLR: Workshop and Conference Proceedings*, 32(1):91–99.

Malhotra, D. and Sebat, J. (2012). CNVs: Harbingers of a Rare Variant Revolution in Psychiatric Genetics.

Marenne, G., Chanock, S. J., Malats, N., and Génin, E. (2013). Advantage of using allele-specific copy numbers when testing for association in regions with common copy number variants. *PloS one*, 8(9):e75350.

Mayrhofer, M., DiLorenzo, S., and Isaksson, A. (2013). Patchwork: allele-specific copy number analysis of whole-genome sequenced tumor tissue. *Genome biology*, 14(3):R24.

McCarroll, S. A., Hadnott, T. N., Perry, G. H., Sabeti, P. C., Zody, M. C., Barrett, J. C., Dallaire, S., Gabriel, S. B., Lee, C., Daly, M. J., and Altshuler, D. M. (2006). Common deletion polymorphisms in the human genome. *Nature genetics*, 38(1):86–92.

McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M. H., de Bakker, P. I. W., Maller, J. B., Kirby, A., Elliott, A. L., Parkin, M., Hubbell, E., Webster, T., Mei, R., Veitch, J., Collins, P. J., Handsaker, R., Lincoln, S., Nizzari, M., Blume, J., Jones, K. W., Rava, R., Daly, M. J., Gabriel, S. B., and Altshuler, D. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature genetics*, 40(10):1166–1174.

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40:4288–4297.

McCullagh, P. (1983). Quasi-Likelihood Functions. *The Annals of Statistics*, 11(1):59–67.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303.

McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., Clouser, C. R., Duncan, C., Ichikawa, J. K., Lee, C. C., Zhang, Z., Ranade, S. S., Dimalanta, E. T., Hyland, F. C., Sokolsky, T. D., Zhang, L., Sheridan, A., Fu, H., Hendrickson, C. L., Li, B., Kotler, L., Stuart, J. R., Malek, J. A., Manning, J. M., Antipova, A. A., Perez, D. S., Moore, M. P., Hayashibara, K. C., Lyons, M. R., Beaudoin, R. E., Coleman, B. E., Laptewicz, M. W., Sannicandro, A. E., Rhodes, M. D., Gottimukkala, R. K., Yang, S., Bafna, V., Bashir, A., MacBride, A., Alkan, C., Kidd, J. M., Eichler, E. E., Reese, M. G., De La Vega, F. M., and Blanchard, A. P. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome research*, 19(9):1527–1541.

Medvedev, P., Fiume, M., Dzamba, M., Smith, T., and Brudno, M. (2010). Detecting copy number variation with mated short reads. *Genome research*, 20(11):1613–22.

Medvedev, P., Stanciu, M., and Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nature methods*, 6(11 Suppl):S13–20.

Meng, X., Saunders, M. A., and Mahoney, M. W. (2014). LSRN: A Parallel Iterative Solver for Strongly Over- or Underdetermined Systems. *SIAM Journal on Scientific Computing*, 36(2):C95–C118.

Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., and Devine, S. E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Research*, 16(9):1182–1190.

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., Abyzov, A., Yoon, S. C., Ye, K., Cheetham, R. K., Chinwalla, A., Conrad, D. F., Fu, Y., Grubert, F., Hajirasouliha, I., Hormozdiari, F., Iakoucheva, L. M., Iqbal, Z., Kang, S., Kidd, J. M., Konkel, M. K., Korn, J., Khurana, E., Kural, D., Lam, H. Y. K., Leng, J., Li, R., Li, Y., Lin, C.-Y., Luo, R., Mu, X. J., Nemesh, J., Peckham, H. E., Rausch, T., Scally, A., Shi, X., Stromberg, M. P., Stütz, A. M.,

Urban, A. E., Walker, J. a., Wu, J., Zhang, Y., Zhang, Z. D., Batzer, M. a., Ding, L., Marth, G. T., McVean, G., Sebat, J., Snyder, M., Wang, J., Ye, K., Eichler, E. E., Gerstein, M. B., Hurles, M. E., Lee, C., McCarroll, S. a., and Korbel, J. O. (2011). Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332):59–65.

Nord, A. S., Lee, M., King, M.-C., and Walsh, T. (2011). Accurate and exact CNV identification from targeted high-throughput sequence data. *BMC genomics*, 12(1):184.

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572.

Park, H., Kim, J.-I., Ju, Y. S., Gokcumen, O., Mills, R. E., Kim, S., Lee, S., Suh, D., Hong, D., Kang, H. P., Yoo, Y. J., Shin, J.-Y., Kim, H.-J., Yavartanoo, M., Chang, Y. W., Ha, J.-S., Chong, W., Hwang, G.-R., Darvishi, K., Kim, H., Yang, S. J., Yang, K.-S., Kim, H., Hurles, M. E., Scherer, S. W., Carter, N. P., Tyler-Smith, C., Lee, C., and Seo, J.-S. (2010). Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nature genetics*, 42(5):400–405.

Plagnol, V., Curtis, J., Epstein, M., Mok, K. Y., Stebbings, E., Grigoriadou, S., Wood, N. W., Hambleton, S., Burns, S. O., Thrasher, A. J., Kumararatne, D., Doffinger, R., and Nejentsev, S. (2012). A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics (Oxford, England)*, 28(21):2747–54.

Pounds, S., Cheng, C., Mullighan, C., Raimondi, S. C., Shurtleff, S., and Downing, J. R. (2009). Reference alignment of SNP microarray signals for copy number analysis of tumors. *Bioinformatics (Oxford, England)*, 25(3):315–21.

Rashid, N. U., Giresi, P. G., Ibrahim, J. G., Sun, W., and Lieb, J. D. (2011a). ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome biology*, 12(7):R67.

Rashid, N. U., Giresi, P. G., Ibrahim, J. G., Sun, W., and Lieb, J. D. (2011b). ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139–140.

Robinson, M. D. and Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23:2881–2887.

Robinson, M. D. and Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9:321–332.

Rokhlin, V. and Tygert, M. (2008). A fast randomized algorithm for overdetermined linear least-squares regression. *Proceedings of the National Academy of Sciences of the United States of America*, 105(36):13212–13217.

Rozowsky, J., Euskirchen, G., Auerbach, R. K., Zhang, Z. D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., and Gerstein, M. B. (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature biotechnology*, 27(1):66–75.

Ruderfer, D. M., Chambert, K., Moran, J., Talkowski, M., Chen, E. S., Gigek, C., Gusella, J. F., Blackwood, D. H., Corvin, A., Gurling, H. M., Hultman, C. M., Kirov, G., Magnusson, P., O'Donovan, M. C., Owen, M. J., Pato, C., St Clair, D., Sullivan, P. F., Purcell, S. M., Sklar, P., and Ernst, C. (2013). Mosaic copy number variation in schizophrenia. *European journal of human genetics : EJHG*, 21(9):1007–11.

Sailani, M. R., Makrythanasis, P., Valsesia, A., Santoni, F. a., Deutsch, S., Popadin, K., Borel, C., Migliavacca, E., Sharp, A. J., Duriaux Sail, G., Falconnet, E., Rabionet, K., Serra-Juhé, C., Vicari, S., Laux, D., Grattau, Y., Dembour, G., Megarbane, A., Touraine, R., Stora, S., Kitsiou, S., Fryssira, H., Chatzisevastou-Loukidou, C., Kanavakis, E., Merla, G., Bonnet, D., Pérez-Jurado, L. a., Estivill, X., Delabar, J. M., and Antonarakis, S. E. (2013). The complex SNP and CNV genetic architecture of the increased risk of congenital heart defects in Down syndrome. *Genome research*, 23(9):1410–21.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., Leotta, A., Pai, D., Zhang, R., Lee, Y.-H., Hicks, J., Spence, S. J., Lee, A. T., Puura, K., Lehtimäki, T., Ledbetter, D., Gregersen, P. K., Bregman, J., Sutcliffe, J. S., Jobanputra, V., Chung, W., Warburton, D., King, M.-C., Skuse, D., Geschwind, D. H., Gilliam, T. C., Ye, K., and Wigler, M. (2007). Strong association of de novo copy number mutations with autism. *Science (New York, N.Y.)*, 316(5823):445–449.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Må nér, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T. C., Trask, B., Patterson, N., Zetterberg, A., and Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. *Science (New York, N.Y.)*, 305(5683):525–528.

Simpson, J. T., McIntyre, R. E., Adams, D. J., and Durbin, R. (2009). Copy number variant detection in inbred strains from short read sequence data. *Bioinformatics*, 26(4):565–567.

Sklar, P., Stone, J. L., ODonovan, M. C., McQuillin, A., Thelander, E. F., Lawrence, J., Holmans, P. A., Kirov, G. K., Medeiros, H., St Clair, D., Owen, M. J., Pato, C. N., Williams, N. M., McGhee, K. A., Puri, V., Milanova, V., Macgregor, S., Sklar (Leader), P., Malloy, P., Lichtenstein, P., Knowles, J. A., Scolnick, E. M., Middleton, F., Ruderfer, D. M., Purcell, S. M., Gill, M., Sullivan, P. F., Curtis, D., Hultman, C. M., Pickard, B., Gates, C., Leh Kwan, S., Williams, H., Conti, D., Crombie, C., Gurling, H., Ardlie, K., McCarroll, S. A., Korn, J., Chambert, K., Walker, N., Daly, M. J., Choudhury, K., Pimm, J., Morris, D. W., Morley, C., Gabriel, S. B., Muir, W. J., Fraser, G., Craddock, N. J., Corvin, A., Blackwood, D. H. R., Georgieva, L., Van Beck, M., Pato, M. T., Kenny, E., Krasucki, R., Helena Azevedo, M., Purcell (Leader), S. M., Maclean, A. W., ODushlaine, C. T., Visscher, P. M., Bass, N., Carvalho, C., Daly, M., Thirumalai, S., Paz Ferreira, C., Datta, S., Mahon, S., Macedo, A., Norton, N., Fanous, A., Waddington, J. L., Toncheva, D., Nikolov, I., Quested, D., and Sullivan, P. (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia.

Stefansson, H., Rujescu, D., Cichon, S., Pietiläinen, O. P. H., Ingason, A., Steinberg, S., Fossdal, R., Sigurdsson, E., Sigmundsson, T., Buizer-Voskamp, J. E., Hansen, T., Jakobsen, K. D., Muglia, P., Francks, C., Matthews, P. M., Gylfason, A., Halldorsson, B. V., Gudbjartsson, D., Thorgeirsson, T. E., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., Bjornsson, A., Mattiasdottir, S., Blondal, T., Haraldsson, M., Magnusdottir, B. B., Giegling, I., Möller, H.-J., Hartmann, A., Shianna, K. V., Ge, D., Need, A. C., Crombie, C., Fraser, G., Walker, N., Lonnqvist, J., Suvisaari, J., Tuulio-Henriksson, A., Paunio, T., Toulopoulou, T., Bramon, E., Di Forti, M., Murray, R., Ruggeri, M., Vassos, E., Tosato, S., Walshe, M., Li, T., Vasilescu, C., Mühleisen, T. W., Wang, A. G., Ullum, H., Djurovic, S., Melle, I., Olesen, J., Kiemeney, L. A., Franke, B., Sabatti, C., Freimer, N. B., Gulcher, J. R., Thorsteinsdottir, U., Kong, A., Andreassen, O. A., Ophoff, R. A., Georgi, A., Rietschel, M., Werge, T., Petursson, H., Goldstein, D. B., Nöthen, M. M., Peltonen, L., Collier, D. A., St Clair, D., and Stefansson, K. (2008). Large recurrent microdeletions associated with schizophrenia. *Nature*, 455(7210):232–236.

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon, R., Bird, C. P., de Grassi, A., Lee, C., Tyler-Smith, C., Carter, N., Scherer, S. W., Tavaré, S., Deloukas, P., Hurles, M. E., and Dermitzakis, E. T. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science (New York, N.Y.)*, 315(5813):848–853.

Sudmant, P. H., Kitzman, J. O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., Bruhn, L., Shendure, J., and Eichler, E. E. (2010). Diversity of human copy number variation and multicopy genes. *Science (New York, N.Y.)*, 330(6004):641–646.

Sullivan, P. F., Daly, M. J., and O'Donovan, M. (2012). Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature reviews. Genetics*, 13(8):537–51.

Sun, W. (2012). A statistical framework for eQTL mapping using RNA-seq data. *Biometrics*, 68(1):1–11.

Sun, W., Wright, F. a., Tang, Z., Nordgard, S. H., Van Loo, P., Yu, T., Kristensen, V. N., and Perou, C. M. (2009). Integrated study of copy number states and genotype calls using high-density SNP arrays. *Nucleic acids research*, 37(16):5365–77.

Szatkiewicz, J. P., Wang, W., Sullivan, P. F., Wang, W., and Sun, W. (2013). Improving detection of copy-number variation by simultaneous bias correction and read-depth segmentation. *Nucleic acids research*, 41(3):1519–32.

Tan, R., Wang, Y., Kleinstein, S. E., Liu, Y., Zhu, X., Guo, H., Jiang, Q., Allen, A. S., and Zhu, M. (2014). An Evaluation of Copy Number Variation Detection Tools from Whole-Exome Sequencing Data. *Human Mutation*, 35(7):899–907.

Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J., and Schork, N. J. (2011). The importance of phase information for human genomics. *Nature reviews. Genetics*, 12(3):215–223.

Treangen, T. J. and Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature reviews. Genetics*, 13(1):36–46.

Tygert, M. (2009). A fast algorithm for computing minimal-norm solutions to underdetermined systems of linear equations. *arXiv preprint arXiv:0905.4745*, 1(3):1–13.

Van Loo, P., Nordgard, S. H., Lingjæ rde, O. C., Russnes, H. G., Rye, I. H., Sun, W., Weigman, V. J., Marynen, P., Zetterberg, A., Naume, B. r., Perou, C. M., Bø rresen Dale, A.-L., and Kristensen, V. N. (2010). Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences of the United States of America*, 107(39):16910–5.

Walters, R. G., Jacquemont, S., Valsesia, A., de Smith, A. J., Martinet, D., Andersson, J., Falchi, M., Chen, F., Andrieux, J., Lobbens, S., Delobel, B., Stutzmann, F., El-Sayed Moustafa, J. S., Chèvre, J.-C., Lecoeur, C., Vatin, V., Bouquillon, S., Buxton, J. L., Boute, O., Holder-Espinasse, M., Cuisset, J.-M., Lemaitre, M.-P., Ambresin, A.-E., Brioschi, A., Gaillard, M., Giusti, V., Fellmann, F., Ferrarini, A., Hadjikhani, N., Campion, D., Guilmatre, A., Goldenberg, A., Calmels, N., Mandel, J.-L., Le Caignec, C., David, A., Isidor, B., Cordier, M.-P., Dupuis-Girod, S., Labalme, A., Sanlaville, D., Béri-Dexheimer, M., Jonveaux, P., Leheup, B., Ounap, K., Bochukova, E. G., Henning, E., Keogh, J., Ellis, R. J., Macdermot, K. D., van Haelst, M. M., Vincent-Delorme, C., Plessis, G., Touraine, R., Philippe, A., Malan, V., Mathieu-Dramard, M., Chiesa, J., Blaumeiser, B., Kooy, R. F., Caiazzo, R., Pigeyre, M., Balkau, B., Sladek, R., Bergmann, S., Mooser, V., Waterworth, D., Reymond, A., Vollenweider, P., Waeber, G., Kurg, A., Palta, P., Esko, T., Metspalu, A., Nelis, M., Elliott, P., Hartikainen, A.-L., McCarthy, M. I., Peltonen, L., Carlsson, L., Jacobson, P., Sjöström, L., Huang, N., Hurles, M. E., O'Rahilly, S., Farooqi, I. S., Männik, K., Jarvelin, M.-R., Pattou, F., Meyre, D., Walley, A. J., Coin, L. J. M., Blakemore, A. I. F., Froguel, P., and Beckmann, J. S. (2010). A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature*, 463(7281):671–675.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F. A., Hakonarson, H., and Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome research*, 17(11):1665–1674.

Wang, Z., Hormozdiari, F., Yang, W.-Y., Halperin, E., and Eskin, E. (2013). CNVeM: copy number variation detection using uncertainty of read mapping. *Journal of computational biology : a journal of computational molecular cell biology*, 20(3):224–36.

Weiss, N., Soules, G., Baum, L. E., and Petrie, T. (1970). A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains.

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G. T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C. L., Irzyk, G. P., Lupski, J. R., Chinault, C., Song, X.-z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D. M., Margulies, M., Weinstock, G. M., Gibbs, R. A., and Rothberg, J. M. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452(7189):872–876.

Wu, J., Grzeda, K. R., Stewart, C., Grubert, F., Urban, A. E., Snyder, M. P., and Marth, G. T. (2012). Copy Number Variation detection from 1000 Genomes project exon capture sequencing data. *BMC bioinformatics*, 13(1):305.

Xi, R., Hadjipanayis, A. G., Luquette, L. J., Kim, T.-M., Lee, E., Zhang, J., Johnson, M. D., Muzny, D. M., Wheeler, D. a., Gibbs, R. a., Kucherlapati, R., and Park, P. J. (2011). Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proceedings of the National Academy of Sciences of the United States of America*, 108(46):E1128–36.

Xie, C. and Tammi, M. T. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC bioinformatics*, 10:80.

Yalcin, B., Wong, K., Agam, A., Goodson, M., Keane, T. M., Gan, X., Nellå ker, C., Goodstadt, L., Nicod, J., Bhomra, A., Hernandez-Pliego, P., Whitley, H., Cleak, J., Dutton, R., Janowitz, D., Mott, R., Adams, D. J., and Flint, J. (2011). Sequence-based characterization of structural variation in the mouse genome. *Nature*, 477(7364):326–329.

Yang, H., Wang, J. R., Didion, J. P., Buus, R. J., Bell, T. A., Welsh, C. E., Bonhomme, F., Yu, A. H.-T., Nachman, M. W., Pialek, J., Tucker, P., Boursot, P., McMillan, L., Churchill, G. A., and de Villena, F. P.-M. (2011). Subspecific origin and haplotype diversity in the laboratory mouse. *Nature genetics*, 43(7):648–655.

Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research*, 19(9):1586–1592.

Zhou, X., Lindsay, H., and Robinson, M. D. (2014). Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Research*, 42.