UCLA
UCLA Electronic Theses and Dissertations

Title
DNA Copy Number Reconstruction via Regularization

Permalink
https://escholarship.org/uc/item/4x08h7p8

Author
Zhang, Zhongyang

Publication Date
2012

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA
Los Angeles

DNA Copy Number Reconstruction via Regularization

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Statistics

by

Zhongyang Zhang

2012

© Copyright by
Zhongyang Zhang
2012

ABSTRACT OF THE DISSERTATION

DNA Copy Number Reconstruction via Regularization

by

Zhongyang Zhang

Doctor of Philosophy in Statistics

University of California, Los Angeles, 2012

Professor Qing Zhou, Co-chair

Professor Chiara Sabatti, Co-chair

Recent advances in genomics have underscored the surprising ubiquity of DNA copy number variants (CNVs). They carry information on the modalities of genome evolution and about the deregulation of DNA replication in cancer cells; their study can be helpful to localize tumor suppressor genes, distinguish different populations of cancerous cell, as well identify genomic variations responsible for disease phenotypes. A number of different high-throughput technologies can be used to identify copy number variable sites, and the literature documents multiple effective algorithms. We augment this literature with a focus on computational speed and simultaneous analysis of multiple sequences.

One the one hand, we explore CNV reconstruction for single sample via estimation with a fused-lasso penalty. We mount a fresh attack on this difficult optimization problem by a majorization-minimization (MM) framework. We also reframe the reconstruction problem in terms of imputation via discrete optimization. This approach is easier and more accurate than parameter estimation. The accuracy of our imputations is comparable to that of hidden Markov models at a substantially lower computational cost.

On the other hand, we investigate the specific problem of detecting regions where

variation in copy number is relatively common in the sample at hand: this encompasses the cases of copy number polymorphisms (CNPs), related samples, technical replicates, and cancerous sub-populations from the same individual. We present a segmentation method to reconstruct CNV regions, that is based on penalized estimation and is capable of processing multiple signals jointly. Our approach is computationally very attractive and leads to sensitivity and specificity levels comparable to those of state-of-the-art specialized methodologies. Its versatility and speed make the method applicable to data obtained with a wide range of technologies and particularly useful in the initial screening stages of large data sets.

Finally, we perform CNV detection and analysis in a set of pedigrees from two Central American isolate and admixed populations. We characterize CNPs in this sample in terms of their frequencies and prevalence on different genetic backgrounds.

The dissertation of Zhongyang Zhang is approved.

Roel A. Ophoff

Yingnian Wu

Kenneth L Lange

Chiara Sabatti, Committee Co-chair

Qing Zhou, Committee Co-chair

University of California, Los Angeles

2012

To the memory of my mother . . .

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I owe everything to my parents for their unconditional love. They always gave me freedom to make my own decisions and pursue my dream from my first school day. They never put any pressure on me, although I know they had high anticipations for me. I can still remember the warm hug of my elder brother right before I get on the flight to US. I could feel his love and support without even a single word.

I am indebted to my teachers at Department of Mathematical Sciences, Tsinghua University. In particular, Professor Yuanlie Lin and Professor Jun Liu introduced me to the fascinating field – statistics.

My days at UCLA were happy and exciting. I feel fortunate to have studied and worked under the guidance of many excellent professors. Professor Chiara Sabatti has been a great advisor. Her invaluable guidance and support had a big impact on me, not only in my study and research, but also in my attitude towards work and life. Professor Kenneth Lange have shaped my mind of what a good statistician and geneticist should possess – rigorous thinking, clear writing, and elegant coding. I would like to thank Professor Yingnian Wu, Professor Qing Zhou, and Professor Roel Ophoff for teaching me and serving on my committee.

My doctorate study would have not been colorful if I did not have so nice a group of fellow students: Janice Brodsky, Mingtian Zhao, Zhangzhang Si, Jean Wang, Xiaofei Yan, Hui Tang, Rui Liu, Hao Wang, Xiaohan Cai, Kuei-Yu Chien, Robert Clements, Miles Chen, Jason Somerville, Rakhee Patel, Yuliya Yaglovskaya, Irina Kukuyeva, Denise Ferrari, Matthew Levinson, Kekona Sorenson, Ka Wong, Fei Fu, Gong Chen, Ryan Rosario, David Zes, etc. Too many to mention you all. Thank you all!

I would like to thank our student affairs officer Glenda Jones, as well as other staff in Department of Statistics and Human Genetics, for their patience and help in administrative issues during my study and research.

I also want to take this opportunity to thank all people for their help and support during my visiting at Stanford University. In particular, Professor Wing Hung Wong gave me the great opportunity to attend his lab meeting and wrote recommendation letter for my job application. Dr. Hui Wang gave me a lot of help and suggestions in my research, life and job search. Professor Charles Lee and his wife Lily Lee and other friends kindly accepted me to a warm and family-like fellowship, which made my life filled with faith and meaning.

Last but most importantly, I am so grateful to my wife, Xuan, for her understanding, support and love.

In my dissertation, Chapter 2 is based on a version of the paper "Z. Zhang, K. Lange, R. Ophoff, and C. Sabatti. Reconstructing DNA copy number by penalized estimation and imputation. The Annals of Applied Statistics, 4(4): 1749-1773, 2010", the principal investigators of which are Dr. Chiara Sabatti and Dr. Kenneth Lange. Chapter 3 is based on a version of the paper "Z. Zhang, K. Lange, and C. Sabatti. Reconstructing DNA copy number by joint segmentation of multiple sequences. Arxiv preprint arXiv:1202.5064, 2012", the principal investigator of which is Dr. Chiara Sabatti.

VITA

2005	B.S., Information and Computing Science Tsinghua University, Beijing, China
2007	M.S., Probability and Mathematical Statistics Tsinghua University, Beijing, China
2009	C.Phil., Statistics University of California, Los Angeles
2008–2012	Graduate Student Researcher Department of Statistics, University of California, Los Angeles
2009–2012	Visiting Student Department of Health Research and Policy, Biostatistics Division, Stanford University

PUBLICATIONS

A. J. Jasinska, S. Service, D. Jawaheer, J. DeYoung, M. Levinson, **Z. Zhang**, B. Kremeyer, H. Muller, I. Aldana, J. Garcia, G. Restrepo, C. Lopez, C. Palacio, C. Duque, M. Parra, J. Vega, D. Ortiz, G. Bedoya, C. Mathews, P. Davanzo, E. Fournier, J. Bejarano, M. Ramirez, C. Araya Ortiz, X. Araya, J. Molina, C. Sabatti, V. Reus, J. Ospina, G. Macaya, A. Ruiz-Linares, N. B. Freimer. A narrow and highly significant linkage signal for severe bipolar disorder in the chromosome 5q33 Region in Latin American pedigrees. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 150B(7): 998-1006, 2009.

Z. Zhang, K. Lange, R. Ophoff, and C. Sabatti. Reconstructing DNA copy number by penalized estimation and imputation. *The Annals of Applied Statistics*, 4(4): 1749-1773, 2010.

A. Ingason, D. Rujescu, S. Cichon, E. Sigurdsson, T. Sigmundsson, O. P. Pietiläinen, J. E. Buizer-Voskamp, E. Strengman, C. Francks, P. Muglia, A. Gylfason, O. Gustafsson, P. I. Olason, S. Steinberg, T. Hansen, K. D. Jakobsen, H. B. Rasmussen, I. Giegling, H. J. Möller, A. Hartmann, C. Crombie, G. Fraser, N. Walker, J. Lonnqvist, J. Suvisaari, A. Tuulio-Henriksson, E. Bramon, L. A. Kiemeney, B. Franke, R. Murray, E. Vassos, T. Toulopoulou, T. W. Mühleisen, S. Tosato, M. Ruggeri, S. Djurovic, O. A. Andreassen, **Z. Zhang**, T. Werge, R. A. Ophoff, GROUP Investigators, M. Rietschel, M. M. Nöthen, H. Petursson, H. Stefansson, L. Peltonen, D. Collier, K. Stefansson, and D. M. St Clair. Copy number variations of chromosome 16p13.1 region associated with schizophrenia. *Molecular Psychiatry* 16(1): 17-25, 2011.

Z. Zhang, K. Lange, and C. Sabatti. Reconstructing DNA copy number by joint segmentation of multiple sequences. *Arxiv preprint arXiv:1202.5064*, 2012.

CHAPTER 1

Introduction

1.1 Biological background

As a more and more comprehensive catalogue of human genetic variation becomes available [The International HapMap Consortium (2005)], a long-term goal in statistical genetics is to investigate the effect of genetic variation, in combination with environment, on human diversity and health, to uncover the mechanism underlying human genome evolution, to understand differences between populations and to identify risk factors of complex diseases and genetic predictors of patient's response to treatment [Goldstein and Cavalleri (2005)].

Variation in DNA happens at a very wide spectrum. Large variants can affect an entire chromosome and typically have detectable phenotypic effect. Down's syndrome, for example, is caused by the presence of an extra copy of Chromosome 21. On the other end of the spectrum, progress in biotechnology allow us to identify sequence-level variation as small as single nucleotide polymorphisms (SNPs), which involve only single base substitution. About 10 million SNPs are estimated to be present in human genome at $> 1\%$ frequency [Feuk et al. (2006)]. They constitute the most common markers used in population genetics and genetic association studies. Between the two extremes, submicroscopic structure variations of intermediate scale are revealed at high resolution with the development of genome-scanning array technologies and comparative DNA sequence analysis.

DNA copy number variants (CNVs) are an important class among other types of

large variants, such as inversion, translocation and loss of heterozygosity (LOH) [Feuk et al. (2006)]. They are generally regarded as gains (insertions or duplications) and losses (deletions or null genotypes) of DNA segments larger than 1kb as compared to their normal counterparts defined by a reference genome or a pool of normal DNA samples [Feuk et al. (2006); Freeman et al. (2006); Scherer et al. (2007)]. In 2004, two landmark papers [Sebat et al. (2004); Iafrate et al. (2004)] opened the door to the investigation of CNVs at the genome-wide scale. Thanks to the steady development of high-throughput technologies, a series of studies [Tuzun et al. (2005); Redon et al. (2006); Jakobsson et al. (2008); Cooper et al. (2008); McCarroll et al. (2008); Kidd et al. (2008); Conrad et al. (2009); The International HapMap 3 Consortium (2010); Mills et al. (2011)] followed and contributed to our increasing knowledge on distributions, sizes, frequencies and other population-genetic properties of CNVs. CNVs occurring in more than 1% of the population are called copy number polymorphisms (CNPs) [Feuk et al. (2006)]. They account for a large proportion of genetic variation in normal cells than were previously expected. The fraction of the genome covered by CNV regions (encompassing overlapping or adjacent gains or losses) is estimated to be 12% based on the populations of the HapMap collection [Redon et al. (2006)]. Similar to SNPs, more than 99% of the CNVs follow Mendelian patterns of inheritance rather than being derived from new mutation [Feuk et al. (2006); McCarroll et al. (2008)]. Most biallelic CNPs are in strong linkage disequilibrium with flanking SNPs [McCarroll et al. (2008); Campbell et al. (2011)], and most low-frequency CNVs segregate on specific SNP haplotypes [McCarroll et al. (2008)]. Given the fact that CNVs cover more proportion of human genome and have higher mutation rate and potentially larger penetrant effect [Sebat (2007); Scherer et al. (2007)], it gradually became a routine to investigate CNVs along with SNPs in a genetic association study.

In the last five years, a number of studies have been carried out that document the significant implications of CNVs in phenotypic variation and complex diseases. Most CNPs are benign and serve as an important resource for human-genetic stud-

ies in diverse worldwide populations. Campbell et al. (2011) reported that biallelic CNPs, in particular those in segmental duplication regions, show greater stratification when compared to frequency-matched SNPs. Several CNPs are associated with susceptibility to common diseases, such as lupus [Fanciulli et al. (2007); Molokhia et al. (2011)], psoriasis [Hollox et al. (2008)], Crohn disease [Bentley et al. (2009)], and obesity [Chen et al. (2011b); Jarick et al. (2011); Yang et al. (2012)]. Recent discoveries have shown that rare and de novo CNVs play a more important role in finding clues to "missing heritability" in some psychiatric diseases, such as autism [Pinto et al. (2010)], schizophrenia [Stefansson et al. (2008); The International Schizophrenia Consortium (2008); Walsh et al. (2008); Vrijenhoek et al. (2008); Ingason et al. (2009)] and bipolar disorder [Malhotra et al. (2011)].

Before the community became aware of their ubiquitous existence in normal cells, changes in copy number were observed in tumor cells. These changes in the cancer genome are mostly due to somatic mutations during the genesis and development of the tumor, so they are usually referred to as copy number aberrations (CNAs) to be distinguishable from inherited CNVs. CNAs are often larger in size and sometimes can extend for an entire chromosome. Recurrent CNAs in a wide range of cancer types have been compiled and cataloged [Albertson et al. (2003)]. It has been shown that duplications in genomic regions harboring oncogenes and deletions in regions containing tumor suppressor genes are involved in the establishment of tumor status through altered gene expression levels [Newton and Lee (2000)], but the general mechanism of complex CNA profiles underlying tumor progression and their evolution in drug resistance is still an open area of research. The study of tumor CNA is complicated by the fact that the available tissue is typically heterogeneous – a mixture of different sub-populations of cancer cells and normal cells [Neuvial et al. (2011)].

Throughout this dissertation, CNV analyses are mainly focused on data generated from normal tissue samples. A simulation study in Chapter 3 is used to demonstrate CNV analysis in tumor data.

1.2 Characteristics of high-throughput technologies

With the continuous advancement of high-throughput biotechnologies, copy number variants can be detected at increasing resolution across the whole genome. Next generation sequencing technology is able to provide a very detailed map of genome-wide structure variants at nucleotide resolution [Mills et al. (2011)], but currently, they are not extensively employed in CNV analysis for large samples due to cost consideration. Microarray based technology remains the primary method used for CNV detection in large-scale genetic studies [Pinto et al. (2011)]. The two main types of microarrays are comparative genomic hybridization (CGH) arrays and single nucleotide polymorphism (SNP) arrays (or genotyping arrays). Characteristics of the data generated by the two platforms will be briefly introduced respectively.

1.2.1 CGH array

CGH array is a high-throughput technology for scanning changes in DNA copy number at genome-wide level [Pinkel et al. (1998); Albertson and Pinkel (2003); Pinkel and Albertson (2005)]. The generated CNV profile falls into a wide range of coverage and resolution depending on the biological materials used as probes distributed on microarrays, such as bacterial artificial chromosomes (BACs) [Snijders et al. (2001)], cDNAs [Pollack et al. (1999)], oligonucleotides [Lucito et al. (2003); Barrett et al. (2004)], but their data generation procedures are of the same nature.

In a typical array-CGH experiment, total genomic DNA is extracted from test and reference cells, fragmented, and differentially labeled with two dyes. The two sets of DNA fragments are then mixed and hybridized to microarrays spotted with probes that are complementary to sample DNA sequences. Fluorescence signals are measured separately for test and reference samples with two different color channels. For a given probe, the relative intensity of the test versus reference signals reflects a noisy measurement of the relative DNA amount, ideally proportional to the relative DNA copy

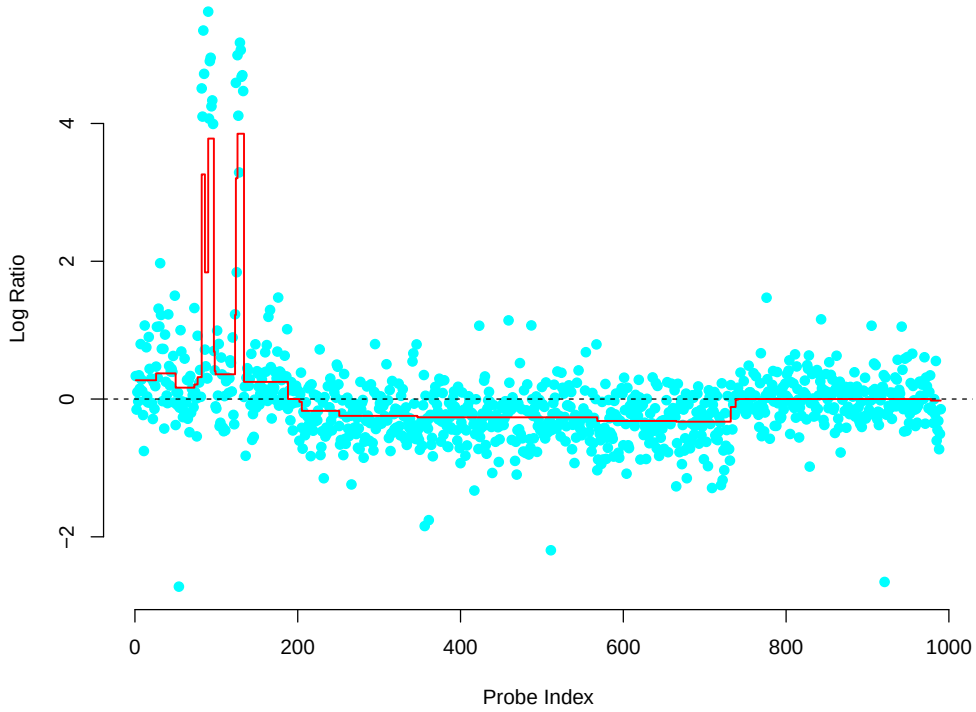


Figure 1.1: An example of tumor array-CGH data. The data is complied in Tibshirani and Wang (2008), showing a genomic region with duplications and deletions in glioblastoma multiforme (GBM) tumors. Log ratio measurements are shown in cyan dots, linearly ordered by probe locations. The red line is fitted by fused lasso [Tibshirani and Wang (2008)]. The dashed line indicates $y = 0$, corresponding to normal copy number.

number in the test versus reference samples at that genomic location [Pinkel and Albertson (2005)]. With proper normalization, the data usually takes the form of log ratio (LogR) of test and reference intensities at each probe, linearly ordered according to the physical locations of probes along the genome.

Figure 1.1 shows an example of array-CGH data present in a genomic region of glioblastoma multiforme (GBM) tumors [Tibshirani and Wang (2008)]. Since variation in DNA copy number tends to occur in contiguous blocks, with probe locations in a block having the same copy number, their log ratio signals appear to fluctuate around a

series of contiguous segments with constant means. Log ratios corresponding to normal copy number are traditionally normalized to 0. The segment with average log intensity ratio larger or smaller than 0 indicate increase or decrease in DNA copy number in the test genome relative to the reference genome.

1.2.2 SNP array

SNP arrays are originally designed for high-density genotyping with hundreds of thousands to millions of probes, but fortunately they also yield information for CNV analysis at no additional cost. Illumina [Peiffer et al. (2006)] and Affymetrix [Bignell et al. (2004); Huang et al. (2004)] are two major commercialized platforms extensively used in scientific community. Despite their obvious technical differences, the two platforms generate conceptually very similar CNV reconstruction problems. For definiteness, we focus on the data delivered by the Illumina platform at our disposal.

When reconstructing CNV from genotype data, researchers rely not only on the final genotype calls but also on raw measurements obtained from the genotyping array. A DNA sample from an individual is preprocessed, hybridized to a chip, and queried at n SNPs. For convenience, we will call the two alleles A and B at each SNP. The amount of DNA carried by each allele at a queried SNP is measured by recording the luminescence of specifically labelled hybridized DNA fragments. Transformations and normalizations of the luminescences lead to two noisy measurements for each SNP i: y_i (LogR or log R ratio (LRR) following Illumina terminology) and x_i (B-allele frequency, BAF). The former quantifies the total DNA present at the SNP location, similar as log intensity ratio generated by CGH array. After normalization, the average of y_i across individuals is 0. A large positive value suggests a duplication; a large negative value suggests a deletion. The distribution of y_i has been successfully described as a mixture of a Gaussian and outliers [Colella et al. (2007); Wang et al. (2009, 2007)].

The B-allele frequency (BAF) represents the fraction of the total DNA attributable

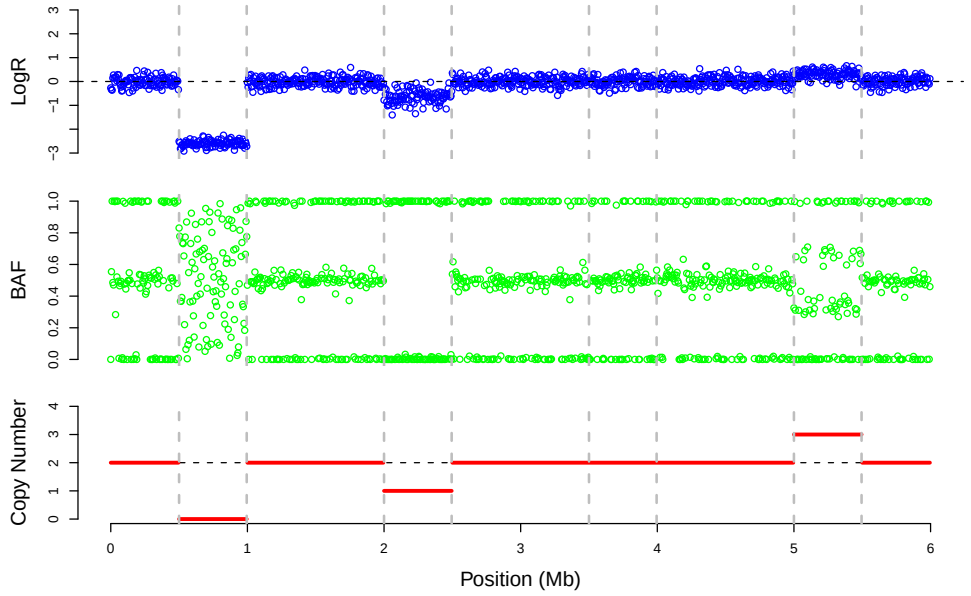


Figure 1.2: Signal patterns for different DNA copy number scenarios organized by their physical locations along a simulated chromosome. The top panel displays in blue y_i (LogR), the middle panel displays in green x_i (BAF), and the bottom panel displays in red the true copy number.

to allele B. The admissible values for x_i occur on the interval $[0, 1]$. When copy number equals 1, x_i takes on values close to 0 or 1, corresponding to the genotypes A and B. When copy number equals 2, x_i is expected to fluctuate around the three possible values 0, 1/2, and 1, corresponding to the three possible genotypes AA, AB, and BB. When copy number equals 3, x_i varies around the four possible values 0, 1/3, 2/3, 1, corresponding to the genotypes AAA, AAB, ABB, BBB. When copy number equals 0, the value of x_i is entirely due to noise and appears to be distributed uniformly on $[0, 1]$. Figure 1.2 plots typical values of the pair (y_i, x_i) along a DNA segment that contains a homozygous deletion (copy number 0), a hemizygous deletion (copy number 1), and a duplication (copy number 3). Clearly both y_i and x_i convey information relevant to copy number.

1.2.3 Data preprocessing

Normalization and transformation of the signal from experimental sources are crucial and can have a very substantial impact on final results. One fundamental assumption in CNV analysis is that markers close together in genomic locations are likely to share the same underlying copy number, so that statistical methods can borrow information across markers to identify chromosomal regions with CNVs. The assumption holds when signals at locus level are properly normalized and some systematic bias due to different artifacts are removed. Some locus-level normalization procedures have been developed to relieve such biases as cross-talk effect between A allele and B allele for microarrays using hybridization, differential probe affinity due to variable PCR fragment length and GC content [Bengtsson et al. (2008, 2009); Ortiz-Estevez et al. (2010)], batch effects arising from laboratory, temporal, or other experimental variation in large studies [Carvalho et al. (2007); Scharpf et al. (2011a,b)]. For tumor CNV data, some heuristic methods have been developed to boost the signal-to-noise ratio at locus level in both settings with matched normal sample [Bengtsson et al. (2010)] or without matched normal sample [Ortiz-Estevez et al. (2012)].

At genome level, a long-range spatial variation in total intensity signals has been widely observed on different platforms including both CGH and SNP arrays. This artifact is highly correlated with local GC content of the DNA sequence and the amplitude of variation is correlated with the deviation of DNA quantity used in experiment from designated amount of standard protocol [Diskin et al. (2008)]. Some fairly intuitive methods have been proposed to eliminate this systematic artifact in practice. Marioni et al. (2007) proposed to fit a loess curve to the original data for each subject and take the residuals as the input for subsequent detection method. This approach has a drawback where signals from some large CNVs are likely to be smoothed out. Instead, Diskin et al. (2008) suggested to regress the total intensity on local GC content with linear model and also use the residuals after removing the fitted linear function. When multiple samples are available, the data can be organized in a matrix form, with rows

corresponding to subjects and columns to queried genomic locations. The artificial pattern shared across samples can be well captured by the first one or two principal components resulting from singular value decomposition (SVD) [Zhang et al. (2010b); Siegmund et al. (2011)].

Indeed, in the data analyses included in the dissertation we need to resort to different data preprocessing strategies and we will describe briefly the fairly standard choices we are making.

1.3 Review of existing methods

Recent genetic studies yield a huge amount of data that can be used for CNV investigation. At the same time, statistical methods and algorithms have been developed to better harness the information available. We refer interested audience to [Lai et al. (2005); Willenbrock and Fridlyand (2005); Dellinger et al. (2010)] for reviews of existing methods in the context of CGH and SNP array data. At the cost of oversimplification, two different approaches have become particularly popular: one is based on the hidden Markov model (HMM) machinery [Rabiner (1989)] and explicitly aims to reconstruct the unobservable DNA copy number; the other, which we will generically call "segmentation", aims at identifying portions of the genome that have constant copy number, without specifically reconstructing it. The pros and cons and their suitable applications are briefly discussed below.

1.3.1 HMM methods

Hidden Markov models and algorithms have dominated the field of CNV reconstruction [Colella et al. (2007); Wang et al. (2007); Korn et al. (2008); Scharpf et al. (2008); Sun et al. (2009); Wang et al. (2009); Yau et al. (2011)], particularly in normal cell data analysis. This statistical framework is flexible enough to accommodate several sources of information, including variable SNP frequencies, variable distances between

adjacent SNPs, linkage disequilibrium, and relationships between study subjects.

In a typical HMM configuration for one subject, assume $\mathbf{o} = (o_1, \ldots, o_N)$ to be the data observed at n marker loci. For CGH array data, $o_i = y_i$, the log intensity ratio of the test sample to the reference sample; whereas for SNP array data, $o_i = (y_i, x_i)$, the bivariate information of total intensity and B-allele frequency. The marker coordinates on the genome $\mathbf{t} = (t_1, \ldots, t_n)$ are also available. The HMM approach takes advantage of the implicitly discrete nature of the copy number process in normal cell data, where the underlying copy number state $s_i \in S$ takes on only a few possible values (6 in PennCNV [Wang et al. (2007)], for example). The transition probability $\mathcal{T}(s_i, s_{i+1}) = f_{s_i \to s_{i+1}}(t_i - t_{i+1})$, as a function of $t_i - t_{i+1}$ (and additional haplotype information), can explicitly account for inter-marker distance (as well as linkage disequilibrium [Wang et al. (2009)]). The emission probabilities $\mathcal{E}(o_i|s_i)$ at each location are assumed to be mutually independent conditional on copy number state s_i. By careful modeling of the emission probabilities, one can fully utilize the information derived from the experimental results. In the case of genotyping arrays, for example, both (y_i, x_i) as well as prior information (for example, minor allele frequencies) can be considered. Some standard algorithms are available for likelihood calculation, inference of hidden states and parameter estimation, including forward and backward recursion procedure, Viterbi's Algorithm, and Baum-Welch algorithm [Rabiner (1989)] (or expectation-minimization (EM) algorithm [Dempster et al. (1977)]).

The HMM approach often has good performance in analysis of normal samples. For example, in a comparison of seven detection methods [Dellinger et al. (2010)] on simulated and real SNP array data, QuantiSNP [Colella et al. (2007)], as a representative of HMM-based methods, outperforms other approaches (including segmentation-based methods) in terms of detection accuracy. With regard to computation, HMM methods are also fast for single sample analysis, because of the limited number of states that need to be included in the search. However, it is non-trivial to extend HMM to the context of multiple sample analysis. For example, a module in PennCNV package is

designed for the joint analysis of a father-mother-child trio to take advantage of relationships between samples [Wang et al. (2008)], but the increase in computational time is expensive to make this analysis impractical for reasonable sample sizes. In the study of cancer, polyploidy and contamination with normal tissues result in a wide range of fractional copy numbers, which violate the discrete nature of copy numbers in normal tissue. Taking these issues into account, some more sophisticated continuous-state HMM methods have been developed for CGH array data [Lai et al. (2008)] and SNP array data [Chen et al. (2011a)]. A significant drawback of these methods is still the heavy load of computation, due to the unconstrained number of hidden states. Possibly for the reasons outlined, HMMs are the methods of choice in the analysis of normal samples.

1.3.2 Segmentation methods

Segmentation based methods for CNV detection resort to change-point models, which itself is an important research topic in statistics. They are based on the assumption that the genome can be partitioned into k segments with each segment harboring constant copy number of DNA. More specifically, the problem is reduced to decide the number of segments k and the location of change points: $1 = \tau_0 < \tau_1 < \ldots < \tau_{k-1} < \tau_k = n$. Choosing k gives rise to a model selection problem while finding the best configuration of change points is a combinatorial problem that involves a searching space with $O(n^{k-1})$ possibilities. To address the former issue, one can use the standard Bayesian information criterion (BIC) [Schwarz (1978)] or a modified version of BIC, specifically designed for CNV analysis [Zhang and Siegmund (2007)]. For a given k, an exact solution to the latter problem can be found at the cost of $O(kn^2)$ in time by dynamic programing [Picard et al. (2005)]. But a problem of such a size is still infeasible provided that the latest generation of CGH and SNP arrays interrogate millions of markers. Methods following two heuristic directions have recently been developed to attack the high-dimensional challenge.

Circular binary segmentation (CBS) [Olshen et al. (2004)] is possibly the most popular change-point detection method in tumor data analysis. It invokes a greedy search that tries to find recursively the best partition of the genome into two or three segments, correspond to single change point or a pair of change points. The stopping criteria is determined by the significance of scan statistics used in partition, which in turn implicitly determines k. A faster version of CBS [Venkatraman and Olshen (2007)] has been implemented using an early stop strategy in p-value calculation, that dramatically reduces the computation cost. In practice, it runs in a reasonable time and performs stably well on different data sets [Lai et al. (2005); Willenbrock and Fridlyand (2005)].

Similar to the idea behind LASSO [Tibshirani (1996)] for model selection in the context of regression, another possible direction counts on relaxing the original non-convex combinatorial problem to a convex optimization problem. Tibshirani and Wang (2008) adapted a model named fused lasso [Tibshirani et al. (2005)] to CNV detection using array CGH data. It aims to solve the optimization problem:

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{i=1}^{N}(y_i-\beta_i)^2+\lambda_1\sum_{i=1}^{N}|\beta_i|+\lambda_2\sum_{i=2}^{N}|\beta_i-\beta_{i-1}|,$$

where y_i is a noisy measure of log intensity ratio and β_i quantifies the true DNA amount at marker i. The method puts an ℓ_1-penalty on jumps between successive markers, which can be viewed as a convex relaxation of ℓ_0 norm, i.e., the number of jumps. The computation cost of the original algorithm is quadratic in n, which can be reduced to linear time [Friedman et al. (2007); Zhang et al. (2010c)]. We will go back to this model and discuss it in more detail in Chapter 2.

An obvious advantage of the segmentation approach is that no a-priori knowledge of the number of copy number states is required and no specific distributional assumption is heavily relied on — making it the standard in cancer studies that involves fractional copy number as discussed above. A limitation of segmentation methods is that they rely on variation in copy number being reflected in the difference in means of the segments—which make them applicable directly to a substantial portion of the data

derived from recent technologies, but not to relative allelic abundance (see the modification suggested in [Staaf et al. (2008)] and following description for an exception). Some additional information like minor allele frequency and linkage disequilibrium is not easy to incorporate. Furthermore, calling procedures that further classify results of segmentation while possibly controlling global error measures [Efron and Zhang (2011)] are also needed.

1.3.3 Multiple sample methods

While a number of successful approaches have been derived along the lines described above for single sample analysis, there is still a paucity of methodology for the joint analysis of multiple sequences. It is clear that if multiple subjects share the same variation in copy number, there exists the potential to increase power by joint analysis. Wang et al. (2009) presented a methodology that extended [Newton and Lee (2000)] to reconstruct the location of tumor suppressor genes from the identification of regions lost in a larger number of samples; the initial steps of the Birdsuite algorithm rely on the identification of suspect signals in the context of multiple samples based on prior knowledge of CNP regions; PennCNV [Wang et al. (2008)] includes an option of joint analysis of trios; methodology to process multiple samples with the context of change point analysis has been developed in a series of papers [Siegmund et al. (2011); Zhang et al. (2010a,b)]; Efron and Zhang (2011) consider FDR analysis of independent samples to identify copy number polymorphysms (CNPs); and Nowak et al. (2011) use a latent feature model to capture, in joint analysis of array-CGH data from multiple tumor samples, shared copy number profiles, on each of which a fused-lasso penalty is enforced for sparsity.

In Chapter 3, we consider a setting similar to [Zhang et al. (2010b)] in that we want joint analysis to inform the segmentation of multiple samples. Our main focus is the analysis of genotyping array data, but the methodology we develop is applicable to a variety of platforms. By adopting a flexible framework we are able, for example,

to define a segmentation algorithm that uses all information from Illumina genotyping data. As in [Siegmund et al. (2011)], we are interested in the situation when not all the samples under consideration carry a copy number variant (CNV): we rather want to enforce a certain sparsity in the vector that identifies which samples carry a given variant. We tackle this problem using a penalized estimation approach, originally proposed in this context by [Tibshirani and Wang (2008)], on which we have developed an algorithmic implementation before [Zhang et al. (2010c)]. Appreciable results are achieved in terms of speed, accuracy and flexibility.

1.4 Summary

Changes in DNA copy number are an important component of genetic variation underlying population diversity and human diseases. Advanced high-throughput technologies, like CGH and SNP arrays, can scan the whole genome for CNVs at high resolution, giving rise to statistical challenges of analyzing a large amount of high-dimensional data. While there has been a number of statistical approaches available for CNV detection with single sample information, the development of joint analysis method of multiple sequences is still worthy further investigation.

The rest of the dissertation is organized as follows: Chapter 2 presents a new algorithm for fused lasso which can be applied to SNP array data efficiently. A discrete version of fused lasso is also proposed, achieving better detection accuracy and computational efficiency. Chapter 3 extends the fused lasso to a more general model called generalized fused lasso. The model can be used to analyze multiple sequences jointly in several CNV applications while maintaining computational efficiency for large data sets. The merits of these methods are demonstrated with both simulated and real data sets. Chapter 4 reports the results of ancestry and CNV analyses on a pedigree data set with samples collected from Costa Rica and Columbia isolates.

CHAPTER 2

Penalized Estimation and Imputation

In this chapter, we investigate the potential of penalized estimation for CNV reconstruction. Tibshirani and Wang (2008) introduced the fused-lasso penalty for the detection of CNVs based on generic considerations of smoothness and sparsity [Rudin et al. (1992); Tibshirani et al. (2005)]. The application of the fused lasso to CNV detection is best motivated by a simplified model. Let the parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_n)$ quantify DNA levels at n successive SNPs. These levels are normalized so that $\beta_i = 0$ corresponds to the standard copy number 2, where SNP i is represented once each on the maternal and paternal chromosomes. Variant regions are rare in the genome and typically involve multiple adjacent SNPs; CNVs range from a few thousand to several million base pairs in length. In high-density genotyping we query SNPs that are on average about a few thousand base pairs apart. The true $\boldsymbol{\beta}$ is therefore expected to be piece-wise constant, with the majority of values equal to 0 and a few segments with positive values (indicating duplication) and negative values (indicating deletion).

Tibshirani and Wang (2008) proposed the joint use of a lasso and a fused-lasso penalty $p(\boldsymbol{\beta}) = \sum_{i=2}^{n} |\beta_i - \beta_{i-1}|$ to enforce this piece-wise constant structure. One then estimates $\boldsymbol{\beta}$ by minimizing the objective function $l(\boldsymbol{\beta}) + \lambda_1 \|\boldsymbol{\beta}\|_{\ell_1} + \lambda_2 p(\boldsymbol{\beta})$, where $l(\boldsymbol{\beta})$ is a goodness-of-fit criteria. The nondifferentiability of the objective function makes minimization challenging [Friedman et al. (2007)]. We mount a fresh attack on this difficult optimization problem by the following tactics: (a) changing penalty terms slightly by substituting a smooth approximation to the absolute value function, (b) majorizing the substitute penalties by quadratics and implementing a new majorization-

minimization (MM) algorithm based on these substitutions, and (c) solving the minimization step of the MM algorithm by a fast version of Newton's method. When the loss function is quadratic, Newton's method takes a single step. More radically, we also reframe the reconstruction problem in terms of imputation via discrete optimization. Readers familiar with Viterbi's algorithm from hidden Markov models will immediately recognize the value of dynamic programming in this context. For the specific problem of detection of CNVs in DNA from normal cells, discrete imputation has the advantage of choosing among a handful of copy number states rather than estimating a continuous parameter. This fact renders discrete imputation easier to implement and more accurate than imputation via parameter estimation.

The chapter is organized as follows. In the methods section, we present our estimation approach to CNV reconstruction and the MM algorithm that implements it. We then describe our new model and the dynamic programming algorithm for discrete imputation. In the results section, we assess the statistical performance and computational speed of the proposed methods on simulated and real datasets.

2.1 Methods

2.1.1 Reconstructing a piece-wise constant function

Consider first CNV reconstruction using signal intensities y_i and neglecting B-allele frequencies x_i. While this restriction overlooks important information, it has the benefit of recasting CNV reconstruction as a general problem of estimating a piecewise constant function from linearly ordered observations. In such regression problems, Tibshirani et al. (2005) and Tibshirani and Wang (2008) suggest minimizing the criterion

$$f(\boldsymbol{\beta}) \;=\; \frac{1}{2}\sum_{i=1}^{n}\Big(y_i-\sum_{j=1}^{p}z_{ij}\beta_j\Big)^2+\lambda_1\sum_{j=1}^{p}|\beta_j|+\lambda_2\sum_{j=2}^{p}|\beta_j-\beta_{j-1}|.$$

Here $\mathbf{y} = (y_i)_{n\times 1}$ is the response vector, $\mathbf{Z} = (z_{ij})_{n\times p}$ is the design matrix, $\boldsymbol{\beta} = (\beta_j)_{n\times 1}$ is the parameter vector of regression coefficients, and λ_1 and λ_2 are tuning parameters that control the sparsity and smoothness of the model. The model is particularly suited to situations where the number of regression coefficients p is much larger than the number of cases n. For the special task of CNV detection, we take $\mathbf{Z} = \mathbf{I}$ (i.e., $p = n$), reducing the objective function to

$$f(\boldsymbol{\beta}) \;=\; \frac{1}{2}\sum_{i=1}^{n}(y_i-\beta_i)^2 + \lambda_1\sum_{i=1}^{n}|\beta_i| + \lambda_2\sum_{i=2}^{n}|\beta_i-\beta_{i-1}|. \tag{2.1}$$

Notice that $f(\boldsymbol{\beta})$ is strictly convex and coercive, so a unique minimum exists. When $\lambda_2 = 0$, the objective function can be decomposed into a sum of n terms, each depending only on one β_i. This makes it very easy to find its minimum using coordinate descent [Friedman et al. (2007); Wu and Lange (2008)]. Unfortunately, this is not the case with $\lambda_2 \neq 0$ because the kinks in the objective function are no longer confined to the coordinate directions. This makes coordinate descent much less attractive [Friedman et al. (2007)]. Quadratic programming [Tibshirani et al. (2005); Tibshirani and Wang (2008)] is still available, but its computational demands do not scale well as p increases.

Inspired by the resolution of similar smoothing dilemmas in imaging [Bioucas-Diaa et al. (2006); Rudin et al. (1992)], we simplify the problem by slightly modifying the penalty. The function

$$||x||_{2,\epsilon} \;=\; \sqrt{x^2+\epsilon}$$

is both differentiable and strictly convex. For small $\epsilon > 0$ it is an excellent approximation to $|x|$. Figure 2.1 illustrates the quality of this approximation for the choice $\epsilon = 0.001$. In practice, we set $\epsilon = 10^{-10}$. If we substitute $||x||_{2,\epsilon}$ for $|x|$, then the CNV objective function becomes

$$f_\epsilon(\boldsymbol{\beta}) \;=\; \frac{1}{2}\sum_{i=1}^{n}(y_i-\beta_i)^2 + \lambda_1\sum_{i=1}^{n}||\beta_i||_{2,\epsilon} + \lambda_2\sum_{i=2}^{n}||\beta_i-\beta_{i-1}||_{2,\epsilon}. \tag{2.2}$$

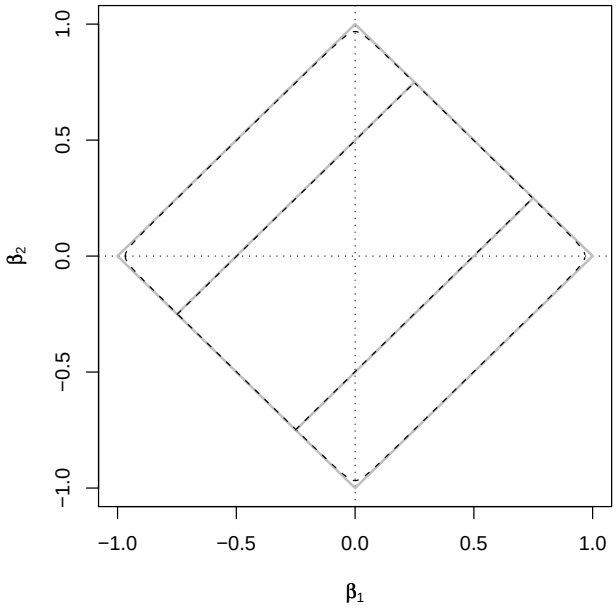


Figure 2.1: Contours corresponding to different penalties. Solid gray line: $|\beta_1|+|\beta_2| = 1$ and $|\beta_1 - \beta_2| = \frac{1}{2}$; Dashed line: $||\beta_1||_{2,\epsilon} + ||\beta_2||_{2,\epsilon} = 1$ and $||\beta_1 - \beta_2||_{2,\epsilon} = \frac{1}{2}$.

As ϵ tends to 0, one can show that the unique minimum point of (2.2) tends to the unique minimum point of the original objective function.

Another virtue of the substitute penalties is that they lend themselves to majorization by a quadratic function. Given the concavity of the function $t \mapsto \sqrt{t+\epsilon}$, it is geometrically obvious that

$$||x||_{2,\epsilon} \quad \leq \quad ||z||_{2,\epsilon} + \frac{1}{2||z||_{2,\epsilon}}[x^2 - z^2],$$

with equality holding if and only if $x = z$. This inequality enables a Majorization-Minimization (MM) [Lange (2004)] strategy that searches for the minimum of the objective function. Each step of this iterative approach requires: (a) majorizing the objective function by a surrogate equal to it at the current parameter vector and (b) minimizing the surrogate. The better-known EM algorithm is a special case of the MM algorithm. The MM algorithm generates a descent path guaranteed to lead to the

optimal solution when one exists. More information can be found in Lange (2004). Returning to our problem, we can replace the objective function by the surrogate function

$$\begin{aligned} g_{\epsilon,m}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(m)}) &= \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_i)^2 \\ &\quad + \frac{\lambda_1}{2}\sum_{i=1}^{n}\frac{\beta_i^2}{||\beta_i^{(m)}||_{2,\epsilon}} + \frac{\lambda_2}{2}\sum_{i=2}^{n}\frac{(\beta_i - \beta_{i-1})^2}{||\beta_i^{(m)} - \beta_{i-1}^{(m)}||_{2,\epsilon}} + c_m, \end{aligned}$$

where m indicates iteration number and c_m is a constant unrelated to $\boldsymbol{\beta}$. Minimization of $g_{\epsilon,m}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(m)})$ to obtain $\boldsymbol{\beta}^{(m+1)}$ drives the objective function $f_\epsilon(\boldsymbol{\beta})$ downhill. Although the MM algorithm entails iteration, it replaces the original problem by a sequence of simple quadratic minimizations. The descent property of the MM algorithm guarantees that progress is made every step along the way. This, coupled with the convexity of our problem, guarantees convergence to the global minimum.

Despite these gains in simplicity, the surrogate function still does not decompose into a sum of n terms, with each depending on only one β_i. The fact that the even numbered β_i do not interact given the odd numbered β_i (and vice versa) suggests alternating updates of the two blocks of even and odd numbered parameters. In practice this block relaxation strategy converges too slowly to be competitive. Fixing β_{i-1} and β_{i+1} leaves too little room to move β_i. Fortunately, full minimization of the quadratic is less onerous than one might expect. The surrogate function can be written in a matrix form

$$g_{\epsilon,m}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(m)}) = \frac{1}{2}\boldsymbol{\beta}^T\mathbf{A}_m\boldsymbol{\beta} - \mathbf{b}_m^T\boldsymbol{\beta} + \tilde{c}_m, \tag{2.3}$$

where $\mathbf{A}_m$ is a tridiagonal symmetric matrix. In view of the strict convexity of the surrogate function, $\mathbf{A}_m$ is also positive definite. The nonzero entries of $\mathbf{A}_m$ and $\mathbf{b}_m$ are

$$a_{1,1}^{(m)} = 1 + \frac{\lambda_1}{||\beta_1^{(m)}||_{2,\epsilon}} + \frac{\lambda_2}{||\beta_2^{(m)} - \beta_1^{(m)}||_{2,\epsilon}};$$

$$a_{i,i}^{(m)} = 1 + \frac{\lambda_1}{||\beta_i^{(m)}||_{2,\epsilon}} + \frac{\lambda_2}{||\beta_i^{(m)} - \beta_{i-1}^{(m)}||_{2,\epsilon}} + \frac{\lambda_2}{||\beta_{i+1}^{(m)} - \beta_i^{(m)}||_{2,\epsilon}}, \quad i = 2, \ldots, n-1;$$

$$
\begin{aligned}
a_{n,n}^{(m)} &= 1 + \frac{\lambda_1}{||\beta_n^{(m)}||_{2,\epsilon}} + \frac{\lambda_2}{||\beta_n^{(m)} - \beta_{n-1}^{(m)}||_{2,\epsilon}}; \\
a_{i,i+1}^{(m)} &= -\frac{\lambda_2}{||\beta_{i+1}^{(m)} - \beta_i^{(m)}||_{2,\epsilon}}, \quad i = 1, \ldots, n-1; \\
a_{i-1,i}^{(m)} &= -\frac{\lambda_2}{||\beta_i^{(m)} - \beta_{i-1}^{(m)}||_{2,\epsilon}}, \quad i = 2, \ldots, n; \\
b_i^{(m)} &= y_i, \quad i = 1, \ldots, n.
\end{aligned}
$$

The minimum of the quadratic occurs at the point $\boldsymbol{\beta} = \mathbf{A}_m^{-1}\mathbf{b}_m$. Thanks to the simple form of $\mathbf{A}_m$, there is a variant of Gaussian elimination known as the tridiagonal matrix algorithm (TDM) or Thomas's algorithm [Conte and deBoor (1972)] that solves the linear system $\mathbf{A}_m\boldsymbol{\beta} = \mathbf{b}_m$ in just $9n$ floating point operations. Alternatively, one can exploit the fact that the Cholesky decomposition of a banded matrix is banded with the same number of bands. As illustrated in Section 2.2.3, Thomas's algorithm is a vast improvement over block relaxation.

A few comments on the outlined strategy are in order. By changing the penalty from $\|\cdot\|_{\ell_1}$ to$||\cdot||_{2,\epsilon}$, we favor less sparse solutions. However, spareness is somewhat besides the point. What we really need are criteria for calling deletions and duplications. The lasso penalty is imposed in this problem because most chromosome regions have a normal copy number where y_i hovers around 0. The same practical outcome can be achieved by imputing copy number 2 for regions where the estimated β_i value is close to 0 (see Section 2.1.3). It is also relevant to compare our minimization strategy to that of [Friedman et al. (2007)]. The fusion step of their algorithm has the advantage of linking coefficients that appear to be similar, but it has the disadvantage that once such links are forged, they cannot be removed. This permanent commitment may preclude finding the global minimum, a limitation that our MM algorithm does not share.

Perhaps more importantly, our strategy can be adapted to handle more general objective functions, as long as the resulting matrix $\mathbf{A}$ in (2.3) is banded, or, at least, sparse. For example, consider the inpainting problem in image reconstruction [Chan and Shen (2002)]. In this two dimensional problem, the intensity levels for certain pixels are lost.

Let S be the set of pixels with known levels. The objective function

$$\begin{aligned} f(\boldsymbol{\beta}) &= \frac{1}{2}\sum_{(i,j)\in S}(y_{ij}-\beta_{ij})^2 \\ &\quad +\lambda\sum_{i=1}^{n}\sum_{j=2}^{n}||\beta_{ij}-\beta_{i,j-1}||_{2,\epsilon}+\lambda\sum_{i=2}^{n}\sum_{j=1}^{n}||\beta_{ij}-\beta_{i-1,j}||_{2,\epsilon} \end{aligned}$$

represents a compromise between imputing unknown values and smoothing. If we majorize the penalties in this objective function by quadratics, then we generate a quadratic surrogate function. The corresponding Hessian of the surrogate is very sparse. (Actually, it is banded, but not in a useful fashion.) Although we can no longer invoke Thomas's algorithm, we can solve the requisite system of linear equations by a sparse conjugate gradient algorithm.

All of the algorithms mentioned so far rely on known values for the tuning constants. We will describe our operational choices for these constants after discussing the problem of imputing chromosome states from estimated parameters in the next section.

2.1.2 Reconstructing discrete copy number states

Imputation of copy number as just described has the drawbacks of neglecting relevant information and requiring the estimation of a large number of parameters. To overcome these limitations, we now bring in the BAF x_i and focus on a model with a finite number of states. This setting brings us much closer to the HMM framework, often used for CNV reconstruction. Such similarity will be evident also in the numerical strategy we will use for optimization. However, our approach avoids the distributional assumptions at the basis of an HMM.

We consider 10 possible genotypic states ϕ, A, B, AA, AB, BB, AAA, AAB, ABB, and BBB at each SNP. Here ϕ is the null state with a copy number of 0. (Note that in the interest of parsimony, we contemplate double deletions, but not double duplications. This has more to do with the strength of signal from duplications than their actual frequency, and it is an assumption that can be easily relaxed.) In the model the average

signal intensity $\mu_{c(s)}$ for a state s depends only on its copy number $c(s)$. Regardless of whether we estimate the μ_c or fix them, they provide a more parsimonious description of the data than the β_i, which could take on a different value for each SNP. Furthermore, while we still need to impute a state for each SNP i, selecting one possible value out of 10 is intrinsically easier than estimation of the continuously varying β_i. Table 2.1 lists the copy number $c(s)$, the expected value of y_i, and the approximate distribution of x_i for each genotype state s. To reconstruct the state vector $\mathbf{s} = (s_1, \ldots, s_n)$, we recommend minimizing the generic objective function

$$\begin{aligned} f(\mathbf{s}) &= \sum_{i=1}^{n} L_1(y_i, s_i) + \alpha \sum_{i=1}^{n} L_2(x_i, s_i) \\ &\quad + \lambda_1 \sum_{i=1}^{n} |\mu_{c(s_i)}| + \lambda_2 \sum_{i=2}^{n} |\mu_{c(s_i)} - \mu_{c(s_{i-1})}|, \end{aligned} \tag{2.4}$$

which again is a linear combination of losses plus penalties. Here α, λ_1, and λ_2 are positive tuning constants controlling the relative influences of the various factors. The lasso penalty makes the states with copy number 2 privileged. The fused-lasso penalty discourages changes in state. Minimizing the objective function (2.4) is a discrete rather than a continuous optimization problem.

Different loss functions may be appropriate in different circumstances. If the intensity values are approximately Gaussian around their means with a common variance, then the choice $L_1(y, s) = [y - \mu_{c(s)}]^2$ is reasonable. For the BAF x_i, the choice $L_2(x, s) = (x - \nu_s)^2$ is also plausible. Here ν_s is the centering constant appearing in the fourth column of Table 2.1. For instance, $L_2(x, \text{ABB}) = (x - 2/3)^2$. For the null state ϕ, we would take

$$L_2(x, \phi) = \int_0^1 (x - u)^2 du = \frac{1}{3}[x^3 + (1 - x)^3].$$

Once the loss functions are set, one can employ dynamic programming to find the state vector $\mathbf{s}$ minimizing the objective function (2.4). If we define the partial solutions

$$g_i(j) = \min_{s_1, \ldots, s_{i-1}} f(s_1, \ldots, s_{i-1}, s_i = j)$$

Table 2.1: Genotype states, corresponding copy numbers, expected values of y_i, and approximate distributions of x_i.

Genotype State s	Copy Number $c(s)$	Mean of y_i	Distribution of x_i
ϕ	0	$\mu_0(<\mu_1)$	Uniform on $[0,1]$
A	1	$\mu_1(<0)$	≈ 0
B	1	$\mu_1(<0)$	≈ 1
AA	2	$\mu_2(\approx 0)$	≈ 0
AB	2	$\mu_2(\approx 0)$	$\approx 1/2$
BB	2	$\mu_2(\approx 0)$	≈ 1
AAA	3	$\mu_3(>0)$	≈ 0
AAB	3	$\mu_3(>0)$	$\approx 1/3$
ABB	3	$\mu_3(>0)$	$\approx 2/3$
BBB	3	$\mu_3(>0)$	≈ 1

for $i = 1, \ldots, n$, then the optimal value of the objective function is $\min_j g_n(j)$. We evaluate the partial solutions $g_i(j)$ recursively via the update

$$\begin{aligned} g_{i+1}(j) &= \min_k \left[g_i(k) + L_1(y_{i+1}, j) + \alpha L_2(x_{i+1}, j) \right. \\ &\quad \left. + \lambda_1 |\mu_{c(j)}| + \lambda_2 |\mu_{c(j)} - \mu_{c(k)}| \right], \end{aligned} \tag{2.5}$$

with initial conditions

$$g_1(j) = L_1(y_1, j) + \alpha L_2(x_1, j) + \lambda_1 |\mu_{c(j)}|.$$

The beauty of dynamic programming is that it applies to a variety of loss and penalty functions.

In fact, it is possible to construct an even more parsimonious model whose four states correspond to the four copy numbers 0, 1, 2, and 3. The loss function $L_1(y, c) = (y - \mu_c)^2$ is still reasonable, but $L_2(x, c)$ should reflect the collapsing of genotypes.

Here c is copy number. Two formulations are particularly persuasive. The first focuses on the minimal loss among the genotypes relevant to each copy number. This produces

$$L_2(x,c) = \begin{cases} \int_0^1 (x-u)^2 du = \frac{1}{3}[x^3 + (1-x)^3], & c = 0, \\ \min\{(x-0)^2, (x-1)^2\}, & c = 1, \\ \min\{(x-0)^2, (x-1/2)^2, (x-1)^2\}, & c = 2, \\ \min\{(x-0)^2, (x-1/3)^2, (x-2/3)^2, (x-1)^2\}, & c = 3. \end{cases} \tag{2.6}$$

The second formulation averages loss weighted by genotype frequency. There are other reasonable loss functions. Among these it is worth mentioning negative log-likelihood, Huber's function, and the hinge loss function of machine learning.

Dynamic programming does require specification of the parameters characterizing the distribution of the intensities y_i and the BAF x_i. It may be possible to assign values to these parameters based on previous data analysis. If not, we suggest estimating them concurrently with assigning states. For example, if the parameters are the intensity means μ_0, μ_1, μ_2, and μ_3, then in practice we alternate two steps starting from plausible initial values for the μ_i. The first step reconstructs the state vector $\boldsymbol{s}$. The second step re-estimates the μ_i conditional on these assignments. Thus, if G_i is the group of SNPs assigned copy number i, then we estimate μ_i by the mean of the y_i over G_i. Taking the median rather the mean makes the process robust to outliers. A few iterations of these two steps usually gives stable parameter estimates and state assignments. To further stabilize the process, we impose two constraints on the second step. If the number of SNPs assigned to G_i is less than a threshold, say 5, we choose not to update μ_i and rather keep the estimate in the previous iteration. In each update we enforce the order of $\mu_0 < \mu_1 < \mu_2 (\approx 0) < \mu_3$. In the following we will refer to the approach described in this section as dynamic programming imputation (DPI).

2.1.3 Identification of deleted and duplicated segments by the fused lasso

In calling deletions and duplications with the fused lasso, we adopt the procedure of [Tibshirani and Wang (2008)]. Originally designed for array-CGH platforms, this pro-

cedure aims to control false discovery rate (FDR). Fortunately, it can be readily applied to genotype data. The general idea is to formulate the problem as one of multiple hypothesis testing for nonoverlapping chromosome segments S_1 through S_K. For each segment S_k we define the test statistic

$$\hat{z}_k \;=\; \frac{\sum_{i\in S_k}\hat{\beta}_i}{\sqrt{n_k}\hat{\sigma}},$$

where n_k is the number of SNPs in segment S_k and $\hat{\sigma}$ is a conservative estimate of standard deviation of the $\hat{\beta}_i$ across all segments based on the y_i values between their 2.5 and 97.5 percentiles. The associated p-value for segment S_k is approximated by $p_k = 2P(Z > |\hat{z}_k|)$ for $Z \sim \mathcal{N}(0,1)$. For a given threshold $q \in (0,1)$, we estimate the FDR by

$$\widehat{\text{FDR}}(q) \;=\; \frac{Kq\cdot\frac{1}{K}\sum_{k=1}^{K} n_k}{\sum_{k=1}^{K} n_k 1_{(p_k\le q)}} \;=\; \frac{q\sum_{k=1}^{K} n_k}{\sum_{k=1}^{K} n_k 1_{(p_k\le q)}}. \tag{2.7}$$

Here the FDR is defined as the ratio between the number of SNPs in nominal CNV segments with true copy number 2 and the total number of SNPs claimed to be within CNV segments. In the FDR estimate (2.7), q is roughly regarded as the fraction of null (copy number 2) segments among all candidate CNV segments. In the numerator, $\frac{1}{K}\sum_{k=1}^{K} n_k$ counts the average SNP number within each segment, and Kq estimates the expected number of null segments. In the denominator, $\sum_{k=1}^{K} n_k 1_{(p_k\le q)}$ counts the total number of SNPs claimed to be located in CNV segments. Thus, this approximation is desired according to the SNP-number-based definition.

Once we decide on an FDR level α, the threshold q is determined as the largest value satisfying $\widehat{\text{FDR}}(q) \le \alpha$. We call a segment S_k a deletion if $\hat{z}_k < 0$ and $p_k \le q$ and a duplication if $\hat{z}_k > 0$ and $p_k \le q$.

2.1.4 Choice of tuning constants

Choice of the tuning constants λ_1 and λ_2 is nontrivial. Because they control the sparsity and smoothness of the parameter vector $\boldsymbol{\beta}$ and therefore drive the process of imputation, it is crucial to make good choices. Both of the references [Friedman et al. (2007)] and

[Wu and Lange (2008)] discuss the problem and suggest solutions in settings similar to ours. While the choice is discussed here in an intuitive way, a theoretical justification for a more general model will be given in Chapter 3.

Friedman et al. (2007) consider the optimal solution to the fused-lasso problem

$$\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2) \;=\; \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_i)^2 + \lambda_1 \sum_{i=1}^{n}|\beta_i| + \lambda_2 \sum_{i=2}^{n}|\beta_i - \beta_{i-1}|.$$

They prove that $\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)$ for $\lambda_1 > 0$ is a soft-thresholding of $\hat{\boldsymbol{\beta}}(0, \lambda_2)$ when $\lambda_1 = 0$, namely,

$$\hat{\beta}_i(\lambda_1, \lambda_2) \;=\; \text{sign}(\hat{\beta}_i(0, \lambda_2))(|\hat{\beta}_i(0, \lambda_2)| - \lambda_1)_+, \quad i = 1, \ldots, n. \tag{2.8}$$

This implies that $\lambda_1 > 0$ will drive to 0 those segments of the piece-wise constant solution $\hat{\boldsymbol{\beta}}(0, \lambda_2)$ whose absolute values are close to 0. It is also important to note that, since $\hat{\boldsymbol{\beta}}(0, \lambda_2)$ is piece-wise constant, its effective dimension is much lower than n.

To understand how the optimal values of these tuning parameters depend on the dimension of the vector $\boldsymbol{\beta}$, let us recall pertinent properties of the Lasso estimator in linear regression. In this setting

$$\hat{\boldsymbol{\beta}} \;=\; \arg\min_{\boldsymbol{\beta}} \frac{1}{2}||\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}||_{\ell_2}^2 + \lambda||\boldsymbol{\beta}||_{\ell_1}, \tag{2.9}$$

where $\mathbf{y}_{n\times 1} \sim \mathcal{N}(\mathbf{Z}_{n\times p}\boldsymbol{\beta}_{p\times 1}, \sigma^2\mathbf{I}_{n\times n})$, and $||\cdot||_{\ell_1}$ and $||\cdot||_{\ell_2}$ are the ℓ_1 and ℓ_2 norms. Candès and Plan (2009), Donoho and Johnstone (1994), and Negahban et al. (2009) show that a Lasso estimator with $\lambda = c\sigma\sqrt{\log p}$ for some constant c leads to an optimal upper bound on $||\mathbf{Z}\boldsymbol{\beta} - \mathbf{Z}\hat{\boldsymbol{\beta}}||_{\ell_2}^2$. Our problem with $\lambda_1 = 0$ fits in this framework if we reparameterize via $\delta_1 = \beta_1$ and $\delta_i = \beta_i - \beta_{i-1}$ for $i = 2, \ldots, n$. In the revised problem

$$\hat{\boldsymbol{\delta}} \;=\; \arg\min_{\boldsymbol{\delta}} \frac{1}{2}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{i}\delta_j)^2 + \lambda_2 \sum_{i=2}^{n}|\delta_i|, \tag{2.10}$$

$p = n$, and the design matrix is lower-triangular with all non-zero entries equal to 1. This finding suggests that we scale λ_2 by $\sqrt{\log n}$.

On the basis of these observations, we explored the choices

$$\lambda_1 \;=\; \rho_1\sigma, \quad \lambda_2 \;=\; \rho_2\sigma\sqrt{\log n}, \quad \rho_1, \rho_2 > 0.$$

Here σ relates the tuning parameters to the noise level. Because the effective dimension in (2.8) is much smaller than n, we assumed that λ_1 does not depend on n. Although ρ_1 and ρ_2 can be tuned more aggressively by cross-validation or criteria such as BIC [Schwarz (1978)] or mBIC [Zhang and Siegmund (2007)], we chose the sensible and operational combination

$$\lambda_1 = \sigma, \quad \lambda_2 \;=\; 2\sigma\sqrt{\log n}. \tag{2.11}$$

A small scale simulation study suggested that the performance of our methods does not vary substantially for values of ρ_1 and ρ_2 close to 1 and 2, respectively. One may also vary ρ_1 and/or ρ_2 mildly to achieve different combinations of sensitivity and specificity as difined in Section 2.2.2 (data not shown).

In practice, we do not know the value of σ. Here we estimated a different σ for each individual, using the standard deviation of y_i values between their 2.5 and 97.5 percentiles. We decided to use only data points within the 95%-interquantile range in order to exclude values of y_i corresponding to possible deletions and duplications. Other possible robust estimators are based on the median absolute deviation or the winsorized standard deviation. In a small-scale simulation we did not observe substantial differences between these estimators (data not shown).

While most of the experiments in the paper used the values of λ_1 and λ_2 suggested in Equation (2.11), we also designed and conducted a more general simulation study to find the optimal values of these tuning parameters; see Section 2.2.6 for details.

2.2 Results

2.2.1 Simulated data with *in silico* CNVs

To illustrate the effectiveness of our algorithms, we tested them on simulated data. We used data from male and female X chromosome to construct *in silico* CNV. Since males are equipped with only one X chromosome, we can use their genotype data to approximate the signal generated by deletion regions. A patchwork of female and male data mimics what we expect from an ordinary pair of homologous chromosomes with occasional deletions. Our X chromosome data come from the schizophrenia study sample of Vrijenhoek et al. (2008) genotyped on the Illumina platform. We focus on the 307 male and 344 female controls.

To avoid artifacts, the data needed to be pre-processed. We identified SNP clusters on the X chromosome using the Beadstudio Illumina software on female controls. These clusters permit estimation of parameters typical of a diploid genome. We then normalized the corresponding male SNP signals relative to the corresponding female signals. Finally, to destroy the signature of possible CNVs in the female data, we permuted the order of the SNPs. This action breaks up the patterns expected within CNV regions and eliminates the smooth variation in the intensity signals [Diskin et al. (2008)].

After these pre-processing steps, we generated ordinary copy number regions from the female data and deleted regions from the male data. We also generated duplications by taking the weighted averages

$$\begin{aligned} y_{i,dup} &= y_{i,f} + 0.55 \times |\text{median}(y_f) - \text{median}(y_m)|, \\ x_{i,dup} &= \frac{1}{3}x_{i,m} + \frac{2}{3}x_{i,f} \end{aligned}$$

for the intensities and BAFs, where the f and m subscripts refer to females and males. Because duplications show a lesser increase in logR values than the deletions show a decrease, the factor 0.55 multiplies the absolute difference $|\text{median}(y_f) - \text{median}(y_m)|$

between median female and male intensities.

We generated two different data sets to assess the operating characteristics of the proposed algorithms. In both data sets the number of deletions equals the number of duplications. *Data set 1* consists of 3600 sequences, each 13000 SNPs long, with either a deletion or a duplication in the central position. The CNVs had lengths evenly distributed over the 6 values 5, 10, 20, 30, 40, and 50 SNPs. *Data set 2* consists of 300 sequences with variable numbers of SNPs and either a deletion or duplication in the central position. The sequence lengths were evenly distributed over the values 4000, 8000, 12000, 16000, and 20000 SNPs; the CNV lengths followed the distribution of data set 1.

The sequence and CNV lengths in our simulations were chosen to roughly mimic values expected in real data. For the Illumina HumanHap550 BeadChip platform, the median number of SNPs per chromosome arm is 13279, with a median absolute deviation of 8172. Current empirical data suggests that there is usually at most one CNV per chromosome arm [Wang et al. (2007)] and that the length of the typical CNV is usually less than 50 SNPs [Jakobsson et al. (2008)]. The sequences from data set 1 represent an average chromosome arm, while the sequences from data set 2 capture the diversity across all chromosome arms. Both data sets have useful lessons to teach.

2.2.2 Measures of accuracy and a benchmark algorithm

We will measure accuracy on a SNP by SNP basis, adopting the following indexes: true positive rate (TPR or sensitivity), false positive rate (FPR or $1-$specificity), and false discovery rate (FDR). These are defined as the ratios

$$\begin{aligned} \text{TPR} &= \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \\ \text{FPR} &= \frac{\text{FP}}{\text{N}} = \frac{\text{FP}}{\text{FP}+\text{TN}}, \\ \text{FDR} &= \frac{\text{FP}}{\text{TP}+\text{FP}}, \end{aligned}$$

where the capital letters T, F, P, N, and R stand for true, false, positive, negative, and rate, respectively. For example, the letter P by itself should be interpreted as the number of SNPs with true copy number equal to 0, 1, or 3; the pair of letters FN should be interpreted as the number of SNPs with true copy number 0, 1, or 3 but imputed copy number 2. We will also evaluate the number of iterations until convergence and the overall computational time required by each algorithm.

For benchmarking purposes, we will compare the performance of the proposed algorithms to that of PennCNV [Wang et al. (2007)], a state-of-the-art hidden Markov model for CNV discovery on Illumina data. PennCNV bases the genotype call for SNP i on its y_i and x_i measurements and its major and minor allele frequencies. We expect PennCNV to perform well because it has been extensively tuned on real and simulated data. The main aim of our comparisons is simply to check whether the new algorithms suffer a substantial loss of accuracy relative to PennCNV.

2.2.3 Convergence of the MMTDM and MMB algorithms

We first investigate two versions of the fused-lasso procedure. Both implement the MM algorithm on the objective function (2.2). The MMTDM algorithm solves the minimization step by the tridiagonal matrix algorithm. The MMB algorithm approximately solves the minimization step by one round of block relaxation. To assess the rate of convergence of MMTDM and MMB, we used data set 1 with 3600 sequences of 13000 SNPs each. We declared convergence for a run when the difference between the objective function at two consecutive iterations fell below 10^{-4}. To limit the computational burden, we set the maximum number of iterations equal to 10000. Both algorithms started with the values $\beta_i = y_i$. Each entry of Table 2.2 summarizes the results for a different CNV width. The table makes it abundantly clear that MMB is not competitive. Because MMB never converged in these trials, we took one sequence and ran it to convergence under the more stringent convergence criterion of 10^{-6}. Figure 2.2 plots the value of the objective function under the two algorithms. Examination of

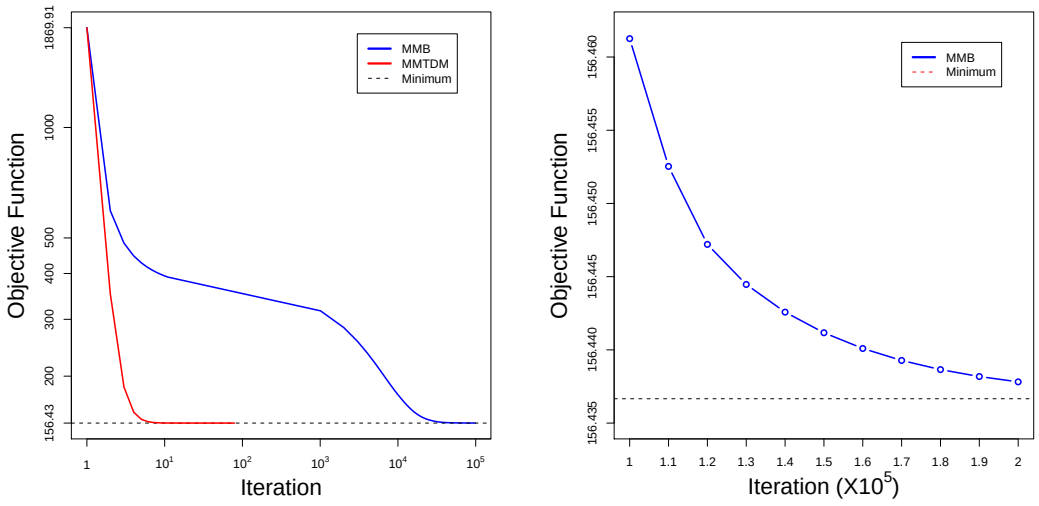
Table 2.2: Number of iterations until convergence of MMTDM and MMB. For MMTDM, each entry summarizes the average number of iterations required for convergence; Standard errors appear in parentheses. MMB never converges within 10000 iterations in this case.

CNV Size	5	10	20	30	40	50
MMB	$>$10000	$>$10000	$>$10000	$>$10000	$>$10000	$>$10000
MMTDM	33.1	33.3	34.5	33.3	33.7	33.9
	(13.0)	(12.0)	(13.9)	(12.9)	(12.2)	(12.1)

these plots shows that MMTDM is on the order of 1000 times faster than MMB.

2.2.4 Effect of including BAF in discrete reconstruction

Data set 1 also illustrates the advantages of including BAF information in CNV reconstruction. Here we focus on dynamic programming imputation (DPI) based on the objective function (2.4). Note that this function does not incorporate prior knowledge of the frequency of deletions versus duplications. In running the dynamic programming algorithm, we rely on results from a previous study [Wang et al. (2009)] to initialize the intensity parameters μ_k. Because the μ_k are re-estimated after each round of imputation, we can safely ignore the slight differences between the genotyping platforms of the previous and current studies. Table 2.3 reports the various accuracy indexes as a function of the tuning constant α determining the relative influence of BAF. Although we already have acceptable reconstruction for $\alpha = 0$, increasing it leads to substantial improvements. When $\alpha = 12$, we reach an excellent balance between sensitivity and specificity. In the following we adopt the value $\alpha = 12$ unless noted to the contrary.



(a) (b)

Figure 2.2: Comparison of convergence rates for the two algorithms MMB and MMTDM for the fused lasso. (a) MMTDM converges much faster than MMB. Blue line: MMB; Red line: MMTDM; Black dashed line: minimum value of objective function; (b) After 10^5 iterations, MMB converges with an accuracy of 0.01.

2.2.5 Accuracy comparisons for various CNV sizes

Table 2.4 reports the values of the accuracy indices for various CNV sizes and types. Here we compare PennCNV, fused-lasso minimization under MMTDM, and DPI on data set 1. To avoid overfitting and a false sense of accuracy, we used 3-fold cross validation to choose α. The accuracy indices reported in the table represent averages over the left-out thirds. Although PennCNV falters a little with the shortest CNVs, it is clearly the best of the three methods. More surprising, DPI achieves comparable FPR and FDR to PennCNV as well as fairly good TPR. In particular, its FDR is uniformly low across CNV sizes and types. Overall, Table 2.4 demonstrates the promise of DPI. In contrast, the results for fused-lasso minimization are discouraging. Despite its post-processing to control FDR, it does poorly in this regard. Furthermore, it displays substantially worse TPR for duplications than PennCNV and DPI, particularly for

Table 2.3: TPR, FPR, and FDR in DPI as α varies.

α	TPR(%)	FPR(%)	FDR(%)	α	TPR(%)	FPR(%)	FDR(%)
0	87.56	0.0064	3.53	15	94.08	0.0010	0.53
1	89.55	0.0031	1.70	16	94.14	0.0010	0.53
2	90.68	0.0019	1.04	17	94.18	0.0010	0.54
3	91.57	0.0017	0.92	18	94.22	0.0011	0.57
4	92.14	0.0014	0.77	19	94.26	0.0011	0.57
5	92.55	0.0012	0.63	20	94.30	0.0012	0.63
6	92.80	0.0010	0.53	21	94.37	0.0012	0.65
7	93.06	0.0010	0.53	22	94.39	0.0013	0.68
8	93.27	0.0010	0.52	23	94.46	0.0015	0.77
9	93.50	0.0010	0.51	24	94.48	0.0015	0.81
10	93.58	0.0009	0.49	25	94.50	0.0016	0.83
11	93.66	0.0009	0.50	26	94.53	0.0016	0.86
12	93.83	0.0009	0.49	27	94.55	0.0018	0.93
13	93.94	0.0009	0.49	28	94.62	0.0018	0.95
14	94.02	0.0010	0.52	29	94.59	0.0019	1.02

duplications spanning only 5 SNPs. This behavior is to be expected given the poor ability of signal strength alone to separate duplications from normal chromosome regions. The performance of fused-lasso minimization underscores the advantages of explicitly modeling the discrete nature of the state space and taking BAF information into account. Nonetheless, it is important to keep in mind that the previous datasets are by design more favorable to PennCNV and DPI. The analysis of tumor samples with ambiguous copy numbers or signals from experimental devices such as CGH arrays that lack allele-specific information are bound to cast fused-lasso minimization in a kinder light. An alternative calling procedure following the fused-lasso segmentation, which

substantially improves the performance of fused lasso, will be discussed in Chapter 3.

2.2.6 Accuracy comparison for various SNP sequence lengths

Data set 2 allowed us to assess performance on longer sequences with less frequent SNPs and to gain insight into the impact of the tuning parameters λ_1 and λ_2. For the latter purpose we adopted two strategies: (a) define λ_1 and λ_2 by the values displayed in Equation (2.11), and (b) adopt an "oracle" approach that relies on the knowledge of locations of deletions and duplications. Strategy (b) chooses constant values across the individuals to maximize TPR (sensitivity) while keeping FPR and FDR levels comparable to those under strategy (a). The oracle approach is not applicable to real data sets, where locations of deletions and duplications are unknown. We adopted it in this analysis to determine how optimal tuning parameters vary with sequence length.

Tables 2.5, 2.6, and 2.7 summarize results for PennCNV, fused-lasso minimization, and DPI, respectively. As with data set 1, PennCNV achieves the best sensitivity, followed by DPI. The best control of false positives occurs with DPI. The accuracy of the methods and the optimal values of λ_1 and λ_2 do not change much with sequence length n, since two extremes $\sqrt{\log(4000)}$ and $\sqrt{\log(20000)}$ as in Equation (2.11) differ slightly. However, it is clear that the advantages of selecting individual-specific λ values outweigh the benefit of selecting constant λ values that maximize overall performance. In fact, the choice of the oracle λ is excessively influenced by some individuals with poor quality data; to control false discoveries in these subjects, one lowers performance in more favorable settings.

2.2.7 Speed comparison of different methods for CNV detection

Finally we compared the computational speeds of the three methods. Although the cost of each scales linearly with the number of SNPs, run times vary considerably in practice (see Figure 2.3). We base our comparisons on data set 2 run on an Intel Xeon

Table 2.4: Accuracy comparison of three methods for various CNV sizes. All accuracy indexes are listed as percentages. The average tuning parameters used in the fused lasso were $\lambda_1 = 0.13(0.04)$ and $\lambda_2 = 0.77(0.22)$; standard deviations appear in parentheses. For DPI, the 3-fold cross validation accuracy indexes are averages over the left-over thirds; initial values of average LogR for each copy number state: $\mu_0 = -5.5923$, $\mu_1 = -0.6313$, $\mu_2 = -0.0045$, $\mu_3 = 0.3252$.

CNV	CNV	PennCNV			Fused Lasso			DPI		
Size	Type	TPR	FPR	FDR	TPR	FPR	FDR	TPR	FPR	FDR
5	Del	83.80	0.0017	4.92	76.67	0.0202	40.66	76.67	0.0006	1.88
	Dup	58.53	0.0011	4.67	00.33	0.0065	98.05	53.60	0.0003	1.28
10	Del	95.03	0.0011	1.45	94.23	0.0130	15.21	89.37	0.0005	0.77
	Dup	93.43	0.0006	0.78	26.00	0.0128	39.01	92.30	0.0006	0.89
20	Del	94.63	0.0008	0.58	96.97	0.0159	09.62	89.87	0.0016	1.15
	Dup	96.13	0.0014	0.92	74.93	0.0126	09.86	95.50	0.0011	0.76
30	Del	94.57	0.0006	0.28	96.76	0.0156	06.53	94.73	0.0013	0.62
	Dup	96.09	0.0001	0.05	85.84	0.0173	08.02	95.39	0.0012	0.55
40	Del	97.83	0.0018	0.59	98.33	0.0158	04.94	98.46	0.0006	0.19
	Dup	94.61	0.0014	0.46	87.88	0.0181	06.24	94.66	0.0012	0.42
50	Del	94.33	0.0003	0.07	95.49	0.0162	04.21	93.82	0.0010	0.26
	Dup	94.50	0.0003	0.09	91.06	0.0121	03.33	95.03	0.0011	0.30
Overall		94.42	0.0009	0.49	88.00	0.0147	07.73	93.70	0.0009	0.50

Table 2.5: Accuracy of PennCNV for various SNP sequence lengths.

Sequence Length	TPR(%)	FPR(%)	FDR(%)
4000	95.54	0.0029	0.46
8000	95.43	0.0019	0.62
12000	96.71	0.0038	1.77
16000	96.46	0.0012	0.74
20000	95.60	0.0007	0.59
Overall	95.95	0.0018	0.84

2.80GHz processor operating under Linux. The PennCNV distributed software (2008, November 19 version) is a combination of C and Perl. We implemented DPI and the MMTDM algorithm for fused-lasso minimization in Fortran 95. The penalty tuning parameters were chosen according to Equation (2.11). For DPI we set $\alpha = 12$. Table 2.8 lists average run times for each sequence sample; standard errors appear in parentheses. As we anticipated, fused-lasso minimization and DPI require less computation per iteration and run much faster than PennCNV. DPI is 2 to 3 times faster than fused-lasso minimization.

2.2.8 Analysis of four real samples

We tested the three methods on genome scan data on four schizophrenia patients from the study of [Vrijenhoek et al. (2008)]. These patients were selected because they each exhibit one experimentally validated CNV (two deletions and two duplications). The four CNVs disrupt the genes MYT1L, CTNND2, NRXN1, and ASTN2, which play important roles in neuronal functioning and are associated with schizophrenia. This subset of the data is ideal for our purpose. The entire data set was collected as part of a genome-wide association study and consists of blood samples from unrelated individuals. It is expected that only a modest amount of CNV may be present; most CNVs

Table 2.6: Accuracy of fused-lasso minimization for various SNP sequence lengths. For strategy (a), average values of λ_1 and λ_2 specified for each individual are summarized for each SNP sequence length; Standard errors appear in parentheses.

Sequence Length	λ_1	λ_2	TPR(%)	FPR(%)	FDR(%)
(a) $\lambda 1$ and $\lambda 2$ specified for each individual according to Equation (2.11)					
4000	0.13 (0.04)	0.73 (0.23)	88.40	0.0414	6.73
8000	0.13 (0.04)	0.76 (0.24)	89.54	0.0241	7.66
12000	0.12 (0.03)	0.76 (0.16)	90.85	0.0148	7.00
16000	0.13 (0.04)	0.79 (0.22)	87.63	0.0103	6.77
20000	0.13 (0.04)	0.80 (0.22)	85.34	0.0084	7.07
Overall	-	-	88.35	0.0145	7.05
(b) Oracle choice of $\lambda 1$ and $\lambda 2$: Constant values across all individuals					
4000	0.16	0.80	83.70	0.0414	7.08
8000	0.19	0.80	77.46	0.0206	7.58
12000	0.18	0.80	84.09	0.0141	7.20
16000	0.17	0.90	81.12	0.0102	7.20
20000	0.18	0.80	76.12	0.0077	7.26
Overall	-	-	80.50	0.0136	7.26

Table 2.7: Accuracy of DPI for various SNP sequence lengths. For strategy (a), average values of λ_1 and λ_2 specified for each individual are summarized for each SNP sequence length; Standard errors appear in parentheses.

Sequence Length	λ_1	λ_2	TPR(%)	FPR(%)	FDR(%)
(a) $\lambda 1$ and $\lambda 2$ specified for each individual according to Equation (2.11)					
4000	0.13 (0.04)	0.73 (0.23)	93.70	0.0013	0.22
8000	0.13 (0.04)	0.76 (0.24)	93.33	0.0007	0.22
12000	0.12 (0.03)	0.76 (0.16)	95.78	0.0004	0.22
16000	0.13 (0.04)	0.79 (0.22)	94.77	0.0009	0.56
20000	0.13 (0.04)	0.80 (0.22)	92.32	0.0005	0.43
Overall	-	-	93.98	0.0007	0.33
(b) Oracle choice of $\lambda 1$ and $\lambda 2$: Constant values across all individuals					
4000	0.15	2.50	87.72	0.0013	0.22
8000	0.24	2.70	86.35	0.0007	0.25
12000	0.12	1.80	94.43	0.0004	0.22
16000	0.18	2.10	91.51	0.0009	0.60
20000	0.16	2.00	90.18	0.0005	0.41
Overall	-	-	90.04	0.0007	0.34

Table 2.8: Computation times for the three CNV imputation methods. The tuning constants in the fused lasso and DPI are noted in Section 2.2.6.

Sequence Length	PennCNV (s)	Fused Lasso (s)	DPI (s)
4000	0.349 (0.034)	0.038 (0.011)	0.011 (0.002)
8000	0.751 (0.111)	0.075 (0.022)	0.023 (0.003)
12000	1.131 (0.145)	0.112 (0.035)	0.057 (0.020)
16000	1.462 (0.181)	0.150 (0.045)	0.077 (0.034)
20000	1.859 (0.260)	0.210 (0.072)	0.099 (0.038)

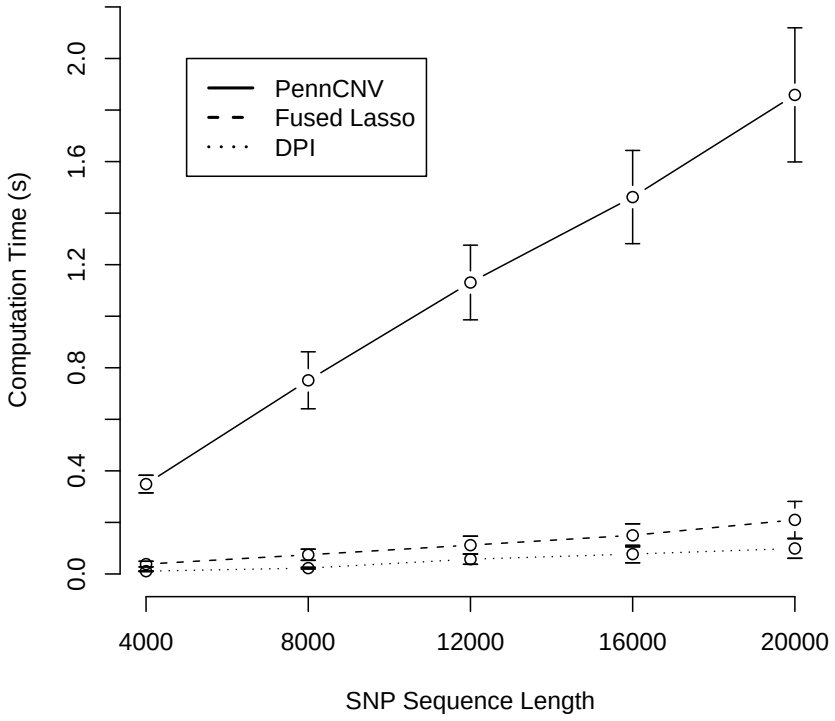


Figure 2.3: Graphical comparison of computation speed as sequence length varies. Solid line: PennCNV; Dashed line: Fused Lasso; Dotted line: DPI.

probably represent inherited neutral polymorphisms rather than de novo mutations. Unlike cancer cell lines, copy numbers should rarely exceed 3.

We analyzed the entire genomes of these four subjects, applying the three methods to each chromosome arm. In calling CNVs with fused-lasso minimization, we controlled FDR at the 0.05 level. The penalty tuning parameters were chosen according to Equation (2.11). For DPI, we set $\alpha = 12$. It took on average 113.8, 8.6, and 4.7 seconds for the three methods to run on the approximately 550k SNPs typed on each individual. The computational efficiency of DPI displayed here may be a decisive advantage in other data sets with thousands of participants. To focus on signals with a higher chance of being real, we eliminated all CNV calls involving fewer than 5 SNPs.

Table 2.9 reports the numbers of detected CNVs and their median sizes; median absolute deviations are listed in parentheses. PennCNV produced the largest number of CNVs calls, followed by fused-lasso minimization. The CNVs detected by PennCNV and DPI had similar sizes; those detected by fused-lasso minimization tended to be longer. Table 2.10 summarizes the overlap between the CNVs calls for the three methods. The vast majority of CNVs detected by DPI are also detected by PennCNV. There is a smaller overlap between PennCNV and the Fused Lasso.

Three of the experimentally verified CNVs were detected by all three methods. The fourth, a deletion on 9q33.1 in patient 4, was detected only by PennCNV (see Figure 2.4). It is noteworthy that the quality of the data for this patient is poor. For example, it fails to pass the PennCNV quality control criterion requiring the standard deviation of LogR to be less than 0.2. In this sample the standard deviation is 0.26. It appears that the higher sensitivity of PennCNV comes at the price of allowing too many false positives. PennCNV calls an exceptionally high number (85) of CNVs for patient 4, with limited overlap with the other two methods.

Table 2.9: CNVs detected by PennCNV, Fused Lasso, and DPI for each patient.

	PennCNV		Fused Lasso		DPI	
Patient	#CNV	CNV Size	#CNV	CNV Size	#CNV	CNV Size
1	34	8 (4)	18	17 (7)	16	10 (4)
2	12	7 (3)	13	11 (9)	7	7 (3)
3	19	8 (4)	14	18 (16)	22	7 (2)
4	85	8 (4)	20	19 (16)	18	9 (4)

2.3 Conclusions

We have proposed two new methods for the reconstruction of CNV. Both methods are much faster than PennCNV, the current state-of-the-art method in CNV discovery. The greater accuracy of DPI versus fused-lasso minimization underscores the importance of using BAF measurements and capitalizing on the discrete nature of CNV imputation. DPI has the additional advantage of outputting the allelic copy numbers so helpful in refining the associations between CNVs and phenotypes. It is hardly surprising that DPI exhibits superior performance in the schizophrenia data where its underlying assumptions hold. By contrast in the analysis of tumor cells, it is much more difficult to fix a priori the number of copies. With its flexibility in fitting piece-wise constant functions to LogR intensities, the fused lasso will shine in this less discrete setting.

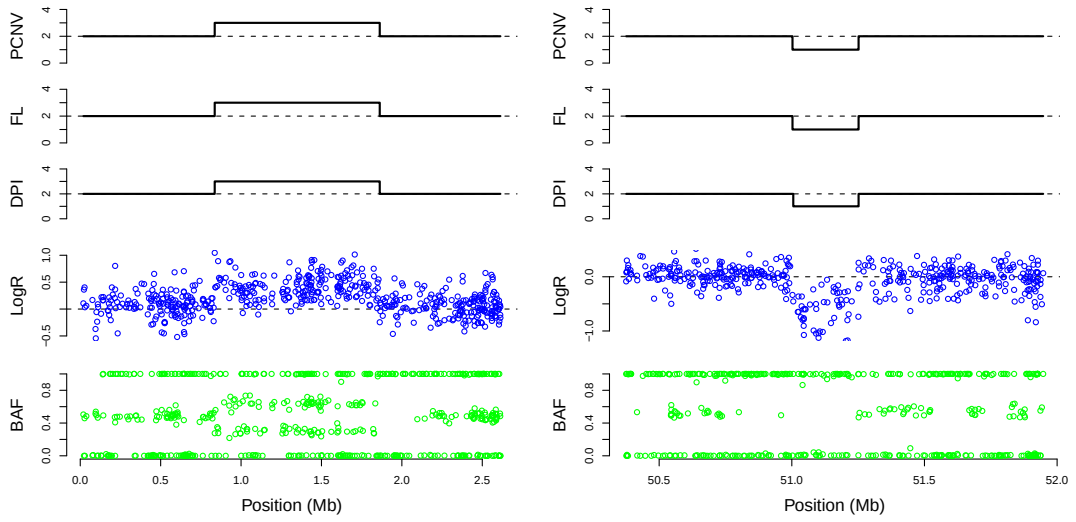
We would like to emphasize that both proposed methods are rough compared to well-established algorithms like PennCNV. There is definitely room for further performance improvements by redefining the loss and penalty functions. As a concrete example, one could modify the fused-lasso penalties to reflect the distances between adjacent SNPs [Li and Zhu (2007)]. We suggest scaling the difference $|\beta_i - \beta_{i-1}|$ by the reciprocal of the physical distance $|t_i - t_{i-1}|$.

In our view penalized models are more parsimonious than hidden Markov models

Table 2.10: Overlap of CNVs detected by PennCNV, Fused Lasso, and DPI. The percentages listed in parentheses refer to the ratio of the number of overlapping CNVs to the total number of unique CNVs detected. For patient 1 DPI treated a large duplication region on the long arm of Chromosome 22 as two segments. Thus, the number of overlapping CNVs was increased by 1 compared to PennCNV vs Fused Lasso.

Patient	PennCNV vs Fused Lasso	PennCNV vs DPI	Fused Lasso vs DPI	3 Methods
1	7 (15.6%)	12 (31.6%)	9 (36.0%)	8 (16.7%)
2	7 (38.9%)	6 (46.2%)	7 (53.8%)	6 (33.3%)
3	10 (43.5%)	15 (57.7%)	8 (28.6%)	8 (26.7%)
4	8 (8.2%)	13 (14.4%)	8 (26.7%)	7 (6.9%)

and achieve many of the same aims. Our redefinition of the fused-lasso penalty and application of the MM algorithm circumvent some of the toughest issues of penalized estimation in the CNV context. In the next chapter, we extend fused lasso to a generalized model that can be applied to joint CNV analysis with multiple samples.



(a) Patient 1: Chr2p25.3

(b) Patient 2: Chr2p16.3

(c) Patient 3: Chr5p15.2

(d) Patient 4: Chr9q33.1

Figure 2.4: PennCNV, fused-lasso minimization, and DPI detected experimentally verified CNVs in 4 schizophrenia patients: (a) A duplication on 2p25.3 of Patient 1; (b) A deletion on 2p16.3 of Patient 2; (c) A duplication on 5p15.2 of Patient 3; (d) A deletion on 9q33.1 of Patient 4. In each subplot from top to bottom, the first three panels display the CNV detected by PennCNV, fused-lasso minimization and DPI respectively, the fourth panel displays in blue y_i (LogR), and the fifth panel displays in green x_i (BAF).

CHAPTER 3

Joint Segmentation of Multiple Sequences

In this chapter, we focus on the specific problem of detecting regions where variation in copy number is relatively common in the sample at hand: this encompasses the cases of copy number polymorphisms, related samples, technical replicates, and cancerous subpopulations from the same individual. We present an algorithm based on regularization approaches with significant computational advantages and competitive accuracy. We illustrate its applicability with simulated and real data sets.

The chapter is organized as follows: Section 3.1 motivates the need for joint analysis of multiple signals. Section 3.2 presents the penalized estimation framework and describes how the model can be used for data analysis by (a) outlining an efficient estimation algorithm, (b) generalizing it to the case of uncoordinated data, and (c) describing the choice of the penalization parameters. Section 3.3 illustrates our results on two simulated data sets (descriptive of normal and tumor samples) and two real data sets: in one case multiple platforms are used to analyze the same sample and in the other case samples from related individuals benefit from joint analysis. Section 3.4 concludes the chapter. Some technique details are deferred to Section 3.5.

3.1 Motivation

The goal of the chapter is to develop a flexible methodology for joint segmentation of multiple sequences that are presumed to carry related information on CNVs. We start by illustrating a series of contexts where the joint analysis appears to be useful.

3.1.1 Genotyping arrays and CNV detection

Genotyping arrays have been used on hundreds of thousands of subjects and the data collected through them provides an extraordinary resource for CNV detection and the study of their frequencies in multiple populations. As detailed in Chapter 1, the raw intensity data representing hybridization strength is processed to obtain two signals: a quantification of total DNA amount (from now on log R Ratio, LRR, following Illumina terminology) and a relative abundance of the two queried alleles (from now on B allele frequency, BAF). Both these signals contain information on CNV and one of the strengths of HMM models has been that they can easily process them jointly. Segmentation models like CBS have traditionally relied only on LRR. While this is a reasonable choice, it can lead to substantial loss of information, particularly in tumor cells, where poliploidy and contamination make information in LRR hard to decipher. To exploit BAF in the context of a segmentation method, a signal transformation has been suggested [Staaf et al. (2008)]: mirrowed BAF (mBAF) relies on exchangeability of the two alleles and the low information content of homozygous SNPs. The resulting mBAF is defined on a coarser grid than the original BAF, but is characterized by changing means in presence of CNV. While Staaf et al. (2008) shows that its analysis alone can be advantageous and more powerful than segmentation of LRR in some contexts, clearly a joint analysis of LRR and mBAF should be preferable to an arbitrary selection of one or the other signal.

3.1.2 Multiple platforms

LRR and BAF are just one example of the multiple signals that one can have available for the same sample. Often, as research progresses, the samples are assessed with a variety of technologies. For example, a number of subjects that have been genotyped at high resolution are now being resequenced. Whenever the technology adopted generates a signal that contains some information on copy number, there is an incentive to

analyze the available signals jointly.

3.1.3 Tumor samples from the same patient obtained at different sites or different progression stages

In an effort to identify mutations that are driving a specific tumor, as well as study its response to treatment, researchers might want to study CNVs in cells obtained at different tumor sites or at different time points [Ostrovnaya et al. (2010)]. Copy number is highly dynamic in cancer cells, so that it is to be expected that some differences be detected over time or across sites. In contrast, the presence of the same CNVs across these samples, can be taken as an indication that the tumors share the same origin: therefore a comparative analysis of CNV can be used to distinguish resurgence of the same cancer from insurgence of a new one, or to identify specific cancer cell populations. Given that the tissue extracted always consists of a mixture of normal and cancer cells, which are in turn a mixture of different populations, joint analysis of the signals from the varied materials is much more likely to lead to the identification of common CNVs, when these exist.

3.1.4 Related subjects

Family data is crucial in genetic investigations and hence it is common to analyze related subjects. When studying individuals from the same pedigree, it is reasonable to assume that some CNVs might be segregating in multiple people: joint analysis would reduce Mendelian errors and increase power of detection.

3.2 Multiple sequence segmentation

3.2.1 A model for joint analysis of multiple signals

Assume we have observed M signals, each measured at N locations, corresponding to ordered physical positions along the genome, with y_{ij} being the observed value of sequence i at location j. Given the nature of the copy number process, we model

$$y_{ij} = \beta_{ij} + \epsilon_{ij}, \tag{3.1}$$

where ϵ_{ij} are i.i.d. noise, and the mean values β_{ij} are piece-wise constant: there exists a linearly ordered partition $\{R_1^{(i)}, R_2^{(i)}, \ldots, R_{K_i}^{(i)}\}$ of the location index $\{1, 2, \ldots, N\}$ such that $\beta_{is} = \cdots = \beta_{it} = \mu_k^{(i)}$ for $s, \ldots, t \in R_k^{(i)}$ and $1 \leq k \leq K_i$. In other words, most of the increments $|\beta_{ij} - \beta_{i,j-1}|$ are assumed to be zero. When two sequences k and l share a CNV with the same boundaries at location j, both $|\beta_{kj} - \beta_{k,j-1}|$ and $|\beta_{lj} - \beta_{l,j-1}|$ will be different from zero in correspondence of the change point. Modulo an appropriate signal normalization, $\beta_{ij} = 0$ can be interpreted as corresponding to the appropriate normal copy number equal to 2. We propose to reconstruct the mean values $\boldsymbol{\beta}$ by minimizing the following function, called hereafter generalized fused lasso (GFL):

$$\begin{aligned} f(\boldsymbol{\beta}) \;=\; & \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N}(y_{ij} - \beta_{ij})^2 + \lambda_1 \sum_{i=1}^{M}\sum_{j=1}^{N}|\beta_{ij}| \\ & + \lambda_2 \sum_{i=1}^{M}\sum_{j=2}^{N}|\beta_{ij} - \beta_{i,j-1}| + \lambda_3 \sum_{j=2}^{N}\left[\sum_{i=1}^{M}(\beta_{ij} - \beta_{i,j-1})^2\right]^{\frac{1}{2}}, \end{aligned} \tag{3.2}$$

which includes a goodness-of-fit term and three penalties, whose roles we will explain one at the time. The ℓ_1 penalty $\sum_{i=1}^{M}\sum_{j=1}^{N}|\beta_{ij}|$ enforces sparsity within $\boldsymbol{\beta}$, in favor of values $\beta_{ij} = 0$, corresponding to the normal copy number. The total variation penalty $\sum_{j=2}^{N}|\beta_{ij} - \beta_{i,j-1}|$ minimizes the number of jumps in the piece-wise constant means of each sequence and was introduced by [Tibshirani and Wang (2008)] in the context of CNV reconstruction from array-CGH data. Finally, the Euclidean penalty on the column vector of jumps $\sqrt{\sum_{i=1}^{M}(\beta_{ij} - \beta_{i,j-1})^2}$ is a form of the group penalty introduced

by [Yuan and Lin (2006)] and favors common jumps across sequences. As clearly explained in [Zhou et al. (2010)], "the local penalty around 0 for each member of a group relaxes as soon as the $|\beta_{ij}-\beta_{i,j-1}|$ for one member i of the group moves off 0." Bleakley and Vert (2011) also suggested the use of this group-fused-lasso penalty to reconstruct CNV. We here consider the use of both the total variation and the Euclidean penalty on the jumps to achieve the equivalent effect of the sparse group lasso, which, as pointed out in [Friedman et al. (2010)], favors CNV detection in multiple samples, allowing for sparsity in the vector indicating which subjects are carriers of the variant. This property is important in situations as presented in Section 3.1.3 and 3.1.4, where one does not want to assume that all the M sequences carry the same CNV.

The incorporation of the latter two penalties can also be naturally interpreted in view of image denoising. To restore an image disturbed by random noise while preserving sharp edges of items in the image, a 2-D total variation penalty $\lambda \sum_{i=1}^{M} \sum_{j=2}^{N} |\beta_{ij} - \beta_{i,j-1}| + \rho \sum_{j=1}^{N} \sum_{i=2}^{M} |\beta_{ij} - \beta_{i-1,j}|$ is proposed in a regularized least-square optimization [Rudin et al. (1992)], where β_{ij} is the true underlying intensity of pixel (i, j). In CNV detection problems, signals from multiple sequences can be aligned up in shape of an image, except that pixels in each sequence are linearly ordered while sequences as a group have no certain order a priori; thus one of the two total variation penalties is replaced by the group penalty on the column vector of jumps.

Using matrix notation, and allowing the tuning parameter λ_1, λ_2 and λ_3 to be sequence specific, we can reformulate the objective function as follows. Let $\mathbf{Y} = (y_{ij})_{M\times N}$ and $\boldsymbol{\beta} = (\beta_{ij})_{M\times N}$. Let $\boldsymbol{\beta}_i$ be the ith row of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_{(j)}$ the jth column of $\boldsymbol{\beta}$. Also, let $\boldsymbol{\lambda}_3 = (\lambda_{3,i})_{M\times 1}$. Then we have

$$\begin{aligned} f(\boldsymbol{\beta}) &= \frac{1}{2}||\mathbf{Y}-\boldsymbol{\beta}||_F^2 + \sum_{i=1}^{M} \lambda_{1,i}||\boldsymbol{\beta}_i||_{\ell_1} \\ &\quad + \sum_{i=1}^{M} \lambda_{2,i}||\boldsymbol{\beta}_{i,2:N} - \boldsymbol{\beta}_{i,1:(N-1)}||_{\ell_1} + \sum_{j=2}^{N} ||\boldsymbol{\lambda}_3 * (\boldsymbol{\beta}_{(j)} - \boldsymbol{\beta}_{(j-1)})||_{\ell_2}, \end{aligned} \tag{3.3}$$

where $||\cdot||_F$ is the Frobenius norm of matrix, $||\cdot||_{\ell_1}$ and $||\cdot||_{\ell_2}$ are ℓ_1 and ℓ_2 norm

of vector, $\boldsymbol{\beta}_{i,s:t}$ indicates the sub-vector with elements $\beta_{i,s}, \ldots, \beta_{i,t}$ in row vector $\boldsymbol{\beta}_i$, and "$*$" is used as entry-wise multiplication between two vectors. Note that it would be easy to modify the tuning parameters so as to make them location specific: that is, reduce the penalty for a jump in correspondence of genomic regions known to harbor CNVs.

3.2.2 An MM algorithm

While the solution to the optimization problem (3.3) might have interesting properties, this approach is useful only if an effective algorithm is available. The last few years have witnessed substantial advances in computational methods for ℓ_1-regularization problems, including the use of coordinate descent [Friedman et al. (2007); Wu and Lange (2008)] and path following methods [Bleakley and Vert (2011); Hoefling (2010); Tibshirani and Taylor (2011); Zhou and Lange (2011)]. The time cost of these methods in the best situation is $O(MNK)$, for K knots along the solution path. It is important to note that these algorithms – some of which are designed for more general applications – may not be the most efficient for large scale CNV analysis for at least two reasons: on the one hand, reasonable choices of λ might be available, making it unnecessary to solve for the entire path; on the other hand, the number of knots K can be expected to be as large as $O(N)$, making the computational costs of path algorithms prohibitive.

With specific regard to the fused-lasso application to CNV detection, we were successful in developing algorithm with per iteration cost $O(N)$ and empirically fast convergence rate for the analysis of one sequence in Chapter 2. We apply the same principles here. We start by modifying the norms in the penalty as follows: rather than the ℓ_1 norm we use $||x||_{2,\epsilon} = \sqrt{x^2 + \epsilon}$ for sufficiently small ϵ, and, for computational stability, we also substitute ℓ_2 norm with $||\mathbf{x}||_{2,\epsilon} = (\sum_{i=1}^{n} x_i^2 + \epsilon)^{\frac{1}{2}}$, obtaining a differentiable

objective function

$$f_\epsilon(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N}(y_{ij}-\beta_{ij})^2 + \sum_{i=1}^{M}\lambda_{1,i}\sum_{j=1}^{N}||\beta_{ij}||_{2,\epsilon} + \sum_{i=1}^{M}\lambda_{2,i}\sum_{j=2}^{N}||\beta_{ij}-\beta_{i,j-1}||_{2,\epsilon} + \sum_{j=2}^{N}||\boldsymbol{\lambda}_3 * (\boldsymbol{\beta}_{(j)}-\boldsymbol{\beta}_{(j-1)})||_{2,\epsilon}. \quad (3.4)$$

Adopting an MM framework [Lange (2004)], we want to find a surrogate function $g_\epsilon(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(m)})$ for each iteration m such that $g_\epsilon(\boldsymbol{\beta}^{(m)} \mid \boldsymbol{\beta}^{(m)}) = f_\epsilon(\boldsymbol{\beta}^{(m)})$ and $g_\epsilon(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(m)}) \geq f_\epsilon(\boldsymbol{\beta})$ for all $\boldsymbol{\beta}$. At each iteration, then, $\boldsymbol{\beta}^{(m+1)} = \arg\min g_\epsilon(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(m)})$. A majorizing function with the above properties is readily obtained using the concavity of square-root function $||x||_{2,\epsilon} \leq \frac{1}{2||z||_{2,\epsilon}}(x^2 - z^2)$, and its vector equivalent $||\mathbf{x}||_{2,\epsilon} \leq \frac{1}{2||\mathbf{z}||_{2,\epsilon}}(||\mathbf{x}||_{\ell_2}^2 - ||\mathbf{z}||_{\ell_2}^2)$. The resulting

$$\begin{aligned} g_\epsilon(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(m)}) = {} & \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N}(y_{ij}-\beta_{ij})^2 + \sum_{i=1}^{M}\lambda_{1,i}\sum_{j=1}^{N}\frac{\beta_{ij}^2}{2||\beta_{ij}^{(m)}||_{2,\epsilon}} \\ & + \sum_{i=1}^{M}\lambda_{2,i}\sum_{j=2}^{N}\frac{(\beta_{ij}-\beta_{i,j-1})^2}{2||\beta_{ij}^{(m)}-\beta_{i,j-1}^{(m)}||_{2,\epsilon}} \\ & + \sum_{j=1}^{N}\frac{||\boldsymbol{\lambda}_3 * (\boldsymbol{\beta}_{(j)}-\boldsymbol{\beta}_{(j-1)})||_{\ell_2}^2}{2||\boldsymbol{\lambda}_3 * (\boldsymbol{\beta}_{(j)}^{(m)}-\boldsymbol{\beta}_{(j-1)}^{(m)})||_{2,\epsilon}} + c^{(m)} \end{aligned}$$

can be decomposed in the sum of similar functions of all the row vectors $\boldsymbol{\beta}_i$

$$g_\epsilon(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(m)}) = \sum_{i=1}^{M} g_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}^{(m)}),$$

where

$$g_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}^{(m)}) = \frac{1}{2}\boldsymbol{\beta}_i\mathbf{A}_i^{(m)}\boldsymbol{\beta}_i^T - [\mathbf{b}_i^{(m)}]^T\boldsymbol{\beta}_i^T + \tilde{c}_i^{(m)}. \quad (3.5)$$

Here each $\mathbf{A}_i^{(m)}$ is a tridiagonal symmetric matrix, and $\tilde{c}_i^{(m)}$ is irrelevant constant for optimization purpose. In view of the strict convexity of the surrogate function, each $\mathbf{A}_i^{(m)}$ is also positive definite. The nonzero entries of $\mathbf{A}_i^{(m)}$ and $\mathbf{b}_i^{(m)}$ $(i = 1, \ldots, M)$ are

listed as follows:

$$
\begin{aligned}
a_i^{(m)}(1,1) &= 1+\frac{\lambda_{1,i}}{||\beta_{i1}^{(m)}||_{2,\epsilon}}+\frac{\lambda_{2,i}}{||\beta_{i2}^{(m)}-\beta_{i1}^{(m)}||_{2,\epsilon}}+\frac{\lambda_{3,i}^2}{||\boldsymbol{\lambda}_3*(\boldsymbol{\beta}_{(2)}^{(m)}-\boldsymbol{\beta}_{(1)}^{(m)})||_{2,\epsilon}};\\
a_i^{(m)}(j,j) &= 1+\frac{\lambda_{1,i}}{||\beta_{ij}^{(m)}||_{2,\epsilon}}+\frac{\lambda_{2,i}}{||\beta_{ij}^{(m)}-\beta_{i,j-1}^{(m)}||_{2,\epsilon}}+\frac{\lambda_{2,i}}{||\beta_{i,j+1}^{(m)}-\beta_{ij}^{(m)}||_{2,\epsilon}}\\
&\quad+\frac{\lambda_{3,i}^2}{||\boldsymbol{\lambda}_3*(\boldsymbol{\beta}_{(j)}^{(m)}-\boldsymbol{\beta}_{(j-1)}^{(m)})||_{2,\epsilon}}+\frac{\lambda_{3,i}^2}{||\boldsymbol{\lambda}_3*(\boldsymbol{\beta}_{(j+1)}^{(m)}-\boldsymbol{\beta}_{(j)}^{(m)})||_{2,\epsilon}},\\
&\qquad\qquad j=2,\ldots,n-1;\\
a_i^{(m)}(n,n) &= 1+\frac{\lambda_{1,i}}{||\beta_{in}^{(m)}||_{2,\epsilon}}+\frac{\lambda_{2,i}}{||\beta_{in}^{(m)}-\beta_{i,n-1}^{(m)}||_{2,\epsilon}}+\frac{\lambda_{3,i}^2}{||\boldsymbol{\lambda}_3*(\boldsymbol{\beta}_{(n)}^{(m)}-\boldsymbol{\beta}_{(n-1)}^{(m)})||_{2,\epsilon}};\\
a_i^{(m)}(j,j-1) &= -\frac{\lambda_{2,i}}{||\beta_{ij}^{(m)}-\beta_{i,j-1}^{(m)}||_{2,\epsilon}}-\frac{\lambda_{3,i}^2}{||\boldsymbol{\lambda}_3*(\boldsymbol{\beta}_{(j)}^{(m)}-\boldsymbol{\beta}_{(j-1)}^{(m)})||_{2,\epsilon}},\\
&\qquad\qquad j=2,\ldots,n;\\
a_i^{(m)}(j,j+1) &= -\frac{\lambda_{2,i}}{||\beta_{i,j+1}^{(m)}-\beta_{ij}^{(m)}||_{2,\epsilon}}-\frac{\lambda_{3,i}^2}{||\boldsymbol{\lambda}_3*(\boldsymbol{\beta}_{(j+1)}^{(m)}-\boldsymbol{\beta}_{(j)}^{(m)})||_{2,\epsilon}},\\
&\qquad\qquad j=1,\ldots,n-1;\\
b_i^{(m)}(j) &= y_{ij},\quad j=1,\ldots,n.
\end{aligned}
$$

Each of the surrogate functions in (3.5) can be minimized solving the linear system $\boldsymbol{\beta}_i=[\boldsymbol{\beta}_i^{(m)}]^T[\mathbf{A}_i^{(m)}]^{-1}$ by the tridiagonal matrix (TDM) algorithm [Conte and deBoor (1972)]. This results in a per-iteraction computational cost of $O(MN)$. This algorithm is empirically observed to achieve an exponential convergence rate (see Section 2.2.3), although we do not yet have an analytic proof. In practice, this method scales well with joint analysis of tens to hundreds of samples with measurements at millions of locations, with limitations dictated by memory requirements. For analysis of real data, we suggest one or a group of samples to be analyzed chromosome by chromosome, since a CNV region can never extend beyond one chromosome to another. Actual computation times are shown along with different examples in Section 3.3.

3.2.3 Stacking observations at different genomic locations

While copy number is continuously defined across the genome, experimental procedures record data at discrete positions, for which we have used the indexes $j = 1, \ldots, N$. In reality, repeated evaluations of the same sample (or related samples) will typically result in measurements at only partially overlapping genomic locations: either because different platforms use different sets of probes, or because missing data my occur at different positions across sequences (consider for example, mBAF and LRR from the same experiment on one subject: the mBAF signal will be defined on a subset of the locations where LRR is).

Let S indicate the union of all genomic positions where some measurement is available among the M signals under study. And let S_i be the subset of locations with measurements in sequence i. We reconstruct β_{ij} for all $j \in S$. When $j \notin S_i$, β_{ij} will be determined simply on the basis of the neighboring data points, relying on the regularizations introduced in (3.3). The goodness-of-fit portion of the objective function is therefore redefined as

$$\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N}(\delta_{ij}y_{ij} - \delta_{ij}\beta_{ij})^2 \quad \text{with} \quad \delta_{ij} = \begin{cases} 1, & \text{if } j \in S_i, \\ 0, & \text{otherwise.} \end{cases} \tag{3.6}$$

The MM strategy previously described applies with slight modifications of the matrix $\mathbf{A}_i^{(m)}$: The item 1 in $a_i^{(m)}(j,j)$ is replaced by δ_{ij} and $b_i^{(m)} = y_{ij}$ is replaced by $b_i^{(m)} = \delta_{ij}y_{ij}$.

The attentive reader would have noted that y_{ij} with $j \notin S_i$ can be considered as missing data, and an evaluation of the characteristics of this missingness is appropriate. In general, y_{ij} cannot be considered missing at random. The most important example is the case of mBAF, where homozygous markers result in missing values. Now, homozygosity is more common when copy number is equal to 1 than when copy number is equal to 2 and, therefore, there is potentially more information on β_{ij} to be extracted from the signals than the one we will capture with the proposed methodology. On the

other hand, it does appear that the approach outlined does not increase false positive: operationally, then, it can be considered as an improvement over segmentation based on LRR only, even if in theory, it does not completely use the information on BAF. It is also relevant to note that, in reality, most of the information on deletion is obtained through LRR, and BAF is really carrying additional information in case of duplications (where the changes in LRR are limited due to saturation effects).

3.2.4 Choice of tuning parameters and segmentation

One of the limitations of penalization procedures is that a value for the tuning parameters needs to be set and clear guidelines are not always available. Path methods that obtain a solution of the optimization problem (3.3) for every value of tuning parameters can be attractive, but recent algorithmic advances [Bleakley and Vert (2011); Tibshirani and Taylor (2011); Zhou and Lange (2011)] remain impractical for problems of the size of ours. A number of recent publications obtain optimal values of penalty parameters under a series of conditions [Bickel et al. (2009); Bunea et al. (2007); Candes and Tao (2007); Donoho and Johnstone (1994)]: we rely upon them to propose the following strategy consisting of obtaining a solution of (3.3) for reasonably liberal values of the tuning parameters, followed by a sequence-by-sequence hard thresholding of the detected jumps with a data-adaptive threshold.

We have found the following guidelines to be useful in choosing penalty parameter values:

$$\begin{aligned} \lambda_{1,i} &= c_1\hat{\sigma}_i, \\ \lambda_{2,i} &= \rho(p)c_2\hat{\sigma}_i\sqrt{\log N}, \\ \lambda_{3,i} &= [1-\rho(p)]c_3\hat{\sigma}_i\sqrt{pM}\sqrt{\log N}, \end{aligned} \tag{3.7}$$

for $i = 1, \ldots, M$, where $\hat{\sigma}_i$ is a robust estimate of standard deviation of $\mathbf{y}_i$, p is roughly the proportion of the M sequences we anticipate to carry CNVs, and c_1, c_2 and c_3 are positive multipliers adjusted in consideration of different signal-to-noise ratios and

CNV sizes.

While a more rigorous justification is provided in Section 3.5, we start by underscoring some of the characteristics of this proposal.

- The sequence-specific penalizing parameters are proportional to an estimate of the standard deviation of the sequence signal: that is, proviso an initial normalization, the same penalties would be used across all signals.

- The tuning parameter for the total variation (fused lasso) and the Euclidean (group fused lasso) penalties on the jumps depend on $\sqrt{\log N}$, where N is the possible number of jumps. This has a "multiple comparison controlling" effect and resembles rates that have been proven optimal under various sparse scenarios [Bickel et al. (2009); Bunea et al. (2007); Candes and Tao (2007); Donoho and Johnstone (1994)]. This term does not appear in the expression of λ_1, as the lasso penalty can be understood as providing a soft thresholding of the solution of (3.3) when $\lambda_1 = 0$: given the penalization due to λ_2 and λ_3, this object will have much smaller dimensionality than N.

- The group penalty depends on $\sqrt{M}$, where M is the number of grouped sequences, as in the original proposal [Yuan and Lin (2006)].

- The relative weight of the fused-lasso and group-fused-lasso penalties is regulated by ρ, which depends on p, the proportion of the M sequences expected to carry the same CNV. For example, if $M = 2$ and the two sequences are LRR and BAF from the same individual, we anticipate $p = 1$ with $\rho = 0$, enforcing jumps at identical places in the two signals. At the other extreme, for completely unrelated sequences, $p = 0$ and $\rho = 1$.

The standard deviation $\hat{\sigma}_i$ can be estimated robustly as follows. Let $\Delta_{ij} = y_{i,j+1} - y_{i,j}$, for $j = 1, \ldots, N-1$, be the one-order difference of adjacent y_{ij} for sequence i.

Then most $\text{Var}(\Delta_{ij}) = 2\sigma_i^2$ except those bridging real change points, so we can take

$$\hat{\sigma}_i = \widehat{SD}(\boldsymbol{\Delta}_i)/\sqrt{2},$$

where

$$\widehat{SD}(\boldsymbol{\Delta}_i) = \text{Standard Deviation}(\boldsymbol{\Delta}_i)$$

or

$$\widehat{SD}(\boldsymbol{\Delta}_i) = \text{Median Absolute Deiviation}(\boldsymbol{\Delta}_i)$$

for $\boldsymbol{\Delta}_i = \{\Delta_{i,1}, \ldots, \Delta_{i,N-1}\}$.

As mentioned before, the exact values of the penalty parameters should be adjusted depending on the expectations of signal strengths. Following the approach in [Rinaldo (2009)], one can approximate the bias induced by each of the penalties and hence work backwards in terms of acceptable levels. As detailed in Section 3.5.1,

$$\begin{aligned}
\text{Bias}(\lambda_1) &\propto \lambda_1; \\
\text{Bias}(\lambda_2) &\propto \lambda_2/\text{Length of segment}; \\
\text{Bias}(\lambda_3) &\propto \lambda_3/(\text{Length of segment} \times \sqrt{\#\text{ sequences sharing segment}}).
\end{aligned}$$

Following again the approach in [Rinaldo (2009)], one can show that under some relatively strong assumptions, the choices in (3.7) lead to a consistent behavior as $N \to \infty$ and M stays bounded (see Section 3.5.2). Despite the fact that N is indeed large in our studies, it is not clear that we can assume it to be in the asymptotic regime. As finer scale measurements become available, scientists desire to investigate CNV of decreasing length: the CNVs we are interested in discovering are often covered by a small number of probes. Furthermore we have often little information on the sizes and frequencies of CNV. In this context, we find it advisable to rely on a two-stage strategy:

1. Sequences are jointly segmented minimizing (3.3) for a relatively lax choice of the penalty parameters.

2. Jumps are further thresholded on the basis of a data-driven cut-off.

Step 2 allows us to be adaptive to the signal strength and can be carried on with multiple methods. For example, one can adopt the modified Bayesian information criteria (mBIC) [Zhang and Siegmund (2007)]. For sequence i, the jumps are sorted as $\{\hat{d}_{i(1)}, \ldots, \hat{d}_{i(N-1)}\}$ in the descending order of their absolute values. And then we choose the first $\hat{k}$ change points where $\hat{k}$ is given by

$$\hat{k} = \arg\max_{k} \text{mBIC}(k).$$

In data analysis, we often apply an even simpler procedure where the threshold for jumps is defined as a fraction of the maximal jump size observed for every sequence. Specifically, for sequence i, let $\hat{D}_i = \max_{2\leq j\leq N}\{|\hat{d}_{ij}|\}$, where $\hat{d}_{ij} = \hat{\beta}_{ij} - \hat{\beta}_{i,j-1}$, be the largest observed jump for sequence i. Then we define

$$\gamma_i = \max\{a\hat{\sigma}_i, \min\{\hat{D}_i, b\hat{\sigma}_i\}\}, \quad \text{for } a < b,$$

as a "ruler" reflecting the scale of a possible real jump size, taking $c\gamma_i$ as the cut-off in removal of most small jumps. In all analyses for this paper, we fix $a = 1$, $b = 5$ and $c = 0.2$. In our experience, this heuristic procedure works well for both tumor and normal tissue CNV data.

3.2.5 Calling procedure

Even if this is not the focus of our proposal, in order to compare the performance of our segmentation algorithm with HMM approaches, it becomes necessary to distinguish acquisitions from losses of copy number. While the same segmentation algorithm can be applied to a wide range of data sets, calling procedures depend more closely on the specific technology used to carry out the experiments. Since our data analysis relies on Illumina genotyping arrays, we limit ourselves to this platform, and briefly describe the calling procedure we adopt in Section 3.3.

Analyzing one subject at the time, each segment with constant mean is assigned to one of five possible copy number states ($c = 0, 1, 2, 3, 4$). Let R collect the indexes of

all SNPs comprising one segment and let $(\mathbf{x}_R, \mathbf{y}_R) = \{(x_j, y_j), j \in R\}$ be the vectors of values for BAF and LRR in the segment. On the basis of typical pattern for BAF and LRR in the different copy number states (see [Colella et al. (2007); Wang et al. (2007, 2009)]), we can write log-likelihood ratio

$$\text{LR}(c) = \log \frac{L_{\text{BAF}}(\mathbf{x}_R; c)}{L_{\text{BAF}}(\mathbf{x}_R; 2)} + \log \frac{L_{\text{LRR}}(\mathbf{y}_R; c)}{L_{\text{LRR}}(\mathbf{y}_R; 2)}, \quad c = 0, 1, 3, 4, \tag{3.8}$$

explicitly defined in Section 3.5.3. Segment R is assigned a CNV state $\hat{c}$ that maximize $\text{LR}(c)$, only if $\text{LR}(\hat{c}) > r_1$, where r_1 is a pre-specified cut-off.

As noted in [Zhang et al. (2010b)], the LRR data for a segment with $c = 2$, ideally normalized to have mean 0, often has a small non-zero mean, due to experimental artifacts. If the number of SNPs in R is sufficiently large, a log-likelihood-ratio criterion as the above would result in the erroneous identification of a copy number different from 2. To avoid this, we also require that the size of the absolute difference of the mean of LRR from zero be larger than a threshold $|\bar{y}_R| > r_2\sigma$.

3.3 Results

We report the results of the analysis of two simulated and two real data sets, which overall exemplify the variety of situations where joint segmentation of multiple sequences is attractive, as described in Section 3.1. In all cases, we compare the performance of the proposed procedure with a set of relevant, often specialized, algorithms. The penalized estimation method we put forward shows competitive performance in all cases and often a substantial computational advantage. Its versatility and speed make it a very convenient tool for initial exploration. To calibrate the run times reported in what follows, it is relevant to know that all our analyses were run on a Mac OS X (10.6.7) machine with 2.93 GHz Intel Core 2 Duo and 4 GB 1067 MHz DDR3 memory.

3.3.1 Simulated CNV in normal samples

We consider one of the simulated data set 1 described in Section 2.2.1. This setting mimics the small rare CNVs possibly occurring in the genome of normal individuals: in our main analysis, therefore, we process one individual at the time, reflecting the typical level of information available to scientists in these contexts. HMM methods, like PennCNV, are expected to be the most effective in this problem; segmentation methods like CBS are closer to our own and therefore also make an interesting comparison. As repeatedly discussed, Illumina platform produces two signals for one subject: LRR and BAF. A segmentation method that can process one signal at the time would give its best results using LRR, which carries most of the information. Given this background, we compare four methods: PennCNV, CBS on LRR, fused lasso on LRR only, and group fused lasso on LRR and mBAF. The implementations we use are those reflected in the software packages: PennCNV (version 2010May01), R package DNAcopy for CBS (version 1.24.0) [Venkatraman and Olshen (2007)] and our own R package Piet (version 0.1.0). Tuning parameters for PennCNV and CBS are set at the default values; the fused lasso implementation corresponds to $\lambda_1 = 0.1$, $\lambda_2 = 2 \times \sqrt{13000}$, and $\lambda_3 = 0$ and the group fused lasso to $\lambda_1 = 0.1$, $\lambda_2 = 0$, and $\lambda_3 = 2 \times \sqrt{13000}$. To call deletion and duplication with CBS and the two fused-lasso approaches, we use both LRR and BAF data (before transformed to mBAF) with the following cut-off values: $r_1 = 10$ and $r_2 = 1(1.5)$ for duplication (deletion). Performance is evaluated by the same indexes we used in Section 2.2.2: true positive rate (TPR or sensitivity) and false discovery rate (FDR), all defined on a per SNP basis. Results are summarized in Table 3.1.

Not surprisingly, all algorithms perform similarly well for larger deletions and duplications and it is mainly for variants that involve $\leq$ 10 SNPs that differences are visible. Algorithms that rely only on LRR (as CBS and fused lasso) underperform in the detection of small duplications (comparison is particularly easy for duplications of size 10 SNP, where the selected parameter values lead to similar FDRs in the three segmentation methods). The group fused lasso can almost entirely recover the performance

Table 3.1: Detection accuracy (as percentage) and computation times for PennCNV, CBS, Fused Lasso and Group Fused Lasso on simulated CNVs in normal samples. Overall accuracy are calculated pooling all sequences with a given type of CNVs. The average (and standard deviation) of the number of seconds required for the analysis of one sequence is reported.

CNV	CNV	PennCNV		CBS		Fused Lasso		Group Fused Lasso	
Size	Type	TPR	FDR	TPR	FDR	TPR	FDR	TPR	FDR
5	Deletion	83.80	4.92	78.20	0.68	63.93	1.74	64.27	1.83
	Duplication	58.53	4.67	11.67	10.26	20.00	37.76	39.87	14.33
10	Deletion	95.03	1.45	88.37	0.56	88.50	0.60	88.87	0.56
	Duplication	93.43	0.78	56.50	4.40	83.90	12.60	91.60	3.85
20	Deletion	94.63	0.58	90.50	0.39	90.80	0.47	90.83	0.47
	Duplication	96.13	0.92	86.22	3.58	92.77	4.95	94.98	2.13
30	Deletion	94.57	0.28	93.30	0.29	89.38	0.52	89.77	0.53
	Duplication	96.09	0.05	90.77	1.61	94.32	1.78	94.98	1.29
40	Deletion	97.83	0.59	97.58	0.09	97.28	0.19	97.28	0.19
	Duplication	94.61	0.46	92.77	0.98	93.94	1.15	94.63	0.75
50	Deletion	94.33	0.07	92.76	0.04	90.47	0.11	90.48	0.11
	Duplication	94.50	0.09	93.81	0.74	93.11	0.79	93.64	0.49
Overall Deletion		95.02	0.55	93.06	0.19	91.08	0.33	91.19	0.34
Overall Duplication		93.82	0.44	86.92	1.55	90.56	2.85	92.46	1.38
Overall		94.42	0.49	89.99	0.85	90.82	1.60	91.83	0.87
Time (sec.)		0.48 (0.01)		0.78 (0.69)		0.22 (0.13)		0.28 (0.05)	

of PennCNV and outperforms CBS in this context.

For curiosity, we analyzed all sequences within each category of CNVs (with the same type and size) simultaneously using GFL. While this represents an unrealistic amount of prior information, it allows us to evaluate the possible gain of joint analysis: FDR practically become 0 ($<$0.02%) for all CNV sizes, but power increases only for CNV including less than 10 SNPs.

Finally, it is useful to compare running times. Summary statistics of the per sample time are reported in Table 3.1: while all algorithms are rather fast, the two implementations of the fused lasso are dominating.

3.3.2 A simulated tumor data set

To explore the challenges presented by tumor data, we rely on a data set created by [Staaf et al. (2008)], with the specific goal of studying the effect of contamination between normal and cancer cells. The HapMap sample NA06991, genotyped on Illumina HumanHap550 array, was used to simulate a cancer cell line, by inserting a total of 10 structure variation regions, including one-copy losses, one-copy gains, and copy neutral loss-of-hetrozygosity (CN-LOH) (see Table 3.2). The signal from this artificial "tumor" sample was then contaminated *in silico* with that of the original "normal" sample, resulting in 21 data sets, with a percentage of normal cells ranging from 0% to 100%. Note that most simulated CNV or CN-LOH regions are very large—some spanning an entire chromosome—and the challenge in detection is really due to the contamination levels.

For ease of comparison, we evaluate the accuracy of calling procedures as in the original reference [Staaf et al. (2008)]: sensitivity is measured for each variant region as the percentage of heterozygous SNPs that are assigned the correct copy number; and specificity is the percentage of originally heterozygous SNPs in unperturbed regions that are assigned CN=2. We compare the performance of GFL to BAFsegmentation

Table 3.2: Regions of allelic imbalance imposed to the HapMap sample NA06991 [Staaf et al. (2008)].

Region	Aberration Type	Chr	bp Start	bp End	#SNP	#hetSNP
1	CN-LOH	5	1	47700000	9397	2756
2	Loss	5	111789971	112521346	156	79
3	Gain	8	1	45200000	12564	3830
4	Gain	8	128432670	129207869	218	91
5	Loss	9	1	50600000	11201	3889
6	Loss	10	84504379	94825178	1988	648
7	Gain	12	1	132449811	27131	8818
8	Loss	13	31766569	31892852	37	10
9	CN-LOH	17	7431864	11747138	1150	308
10	CN-LOH	17	22300000	78774742	9713	3205
Total number of modified heterozygous SNPs						23634
Total number of heterozygous SNPs on autosome						176207
Total number of SNPs on autosome						547359

[Staaf et al. (2008)] and PSCN [Chen et al. (2011a)] representing, respectively, a version of segmentation and HMM approaches specifically developed to deal with contaminated tumor samples (both these algorithms have been tested with success on this simulated data set).

Following other analyses, we do not pre-process the data prior to CNV detection. BAFsegmentation and PSCN were run using recommended parameter values. For each of the diluted data sets, we applied the GFL model on each chromosome at one time using both LRR and mBAF, whose standard deviations are normalized to 1. Tuning constants are set to $\lambda_1 = 0$, $\lambda_2 = 0.5 \times 3 \times \sqrt{\log N}$, and $\lambda_3 = 0.5 \times 3 \times \sqrt{\log N}$, varying specifically for chromosome interrogated by N SNPs. The change points resulting from hard segmentation on LRR and mBAF are combined to make a finer segmentation of the genome. Finally, we adopt the same calling procedure described by [Staaf et al. (2008)]. For ease of comparison with PSCN, only analysis of simulated tumor data are reported, even if BAFsegmentation and GFL would gain from using the genotype of normal cell in defining mBAF.

Figure 3.1 summarizes the sensitivity of each method, as a function of percentage of normal cell in the sample. Sensitivity is calculated for each of the 10 regions separately. All three methods work reasonably well under a wide range of percentages of normal cell contamination (in 5 out of the 10 regions, GFL appears to lead to best results, while in the other 5 PSCN does). The CNV region that comprises the smallest amount of SNP is the hemizygous loss on Chromosome 13: in this case GFL in our hands behaved in the most stable manner. GLF outperforms the two comparison methods in terms of specificity (Figure 3.2): while the specificity values might appear very high in any case, this is somewhat of an artifact due to the adopted definition of this index. It is relevant to note that the performance of PSCN in our hands does not correspond to the published one [Chen et al. (2011a)]. While we tried our best to set the parameter values, we have not succeeded in replicating the authors' original results, which should be considered in the interest of fairness.

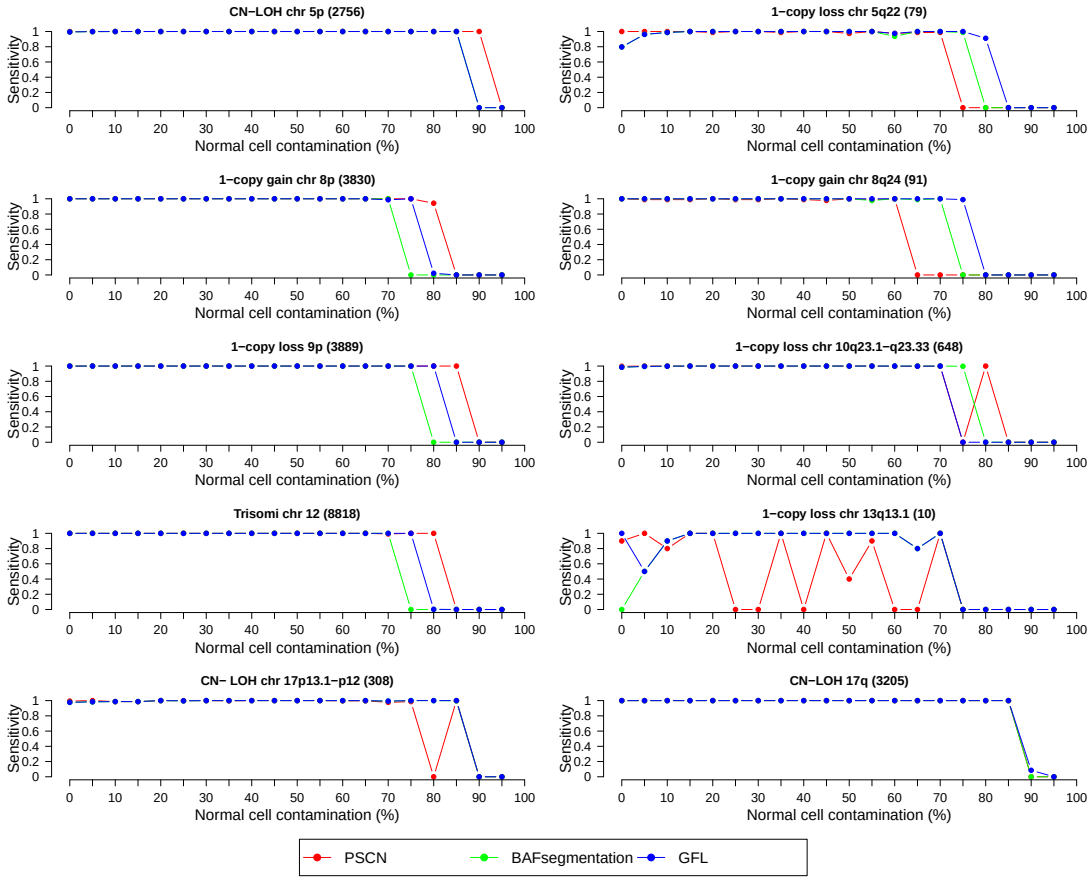


Figure 3.1: Sensitivity as function of percentage contamination by normal cells in the 10 different simulated CNV regions. Sensitivity is not defined at 100% contamination.

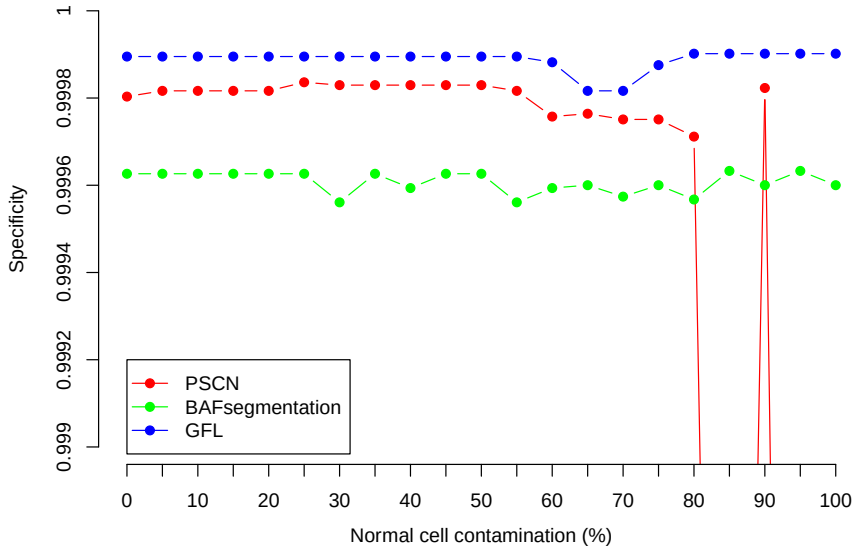


Figure 3.2: Specificity as function of percentage contamination by normal cells. Note that Chen et al. (2011a) reports better perfomance of PSCN in correspondence of contamination levels 85% , 95% and 100%.

PSCN, like GFL, is implemented in R with some computationally intensive subroutines coded in C. BAFsegmentation relies its segmentation part on the R package DNAcopy, whose core algorithms are implemented in C and Fortran, and it is wrapped in Perl. A comparison of run times indicate that GLF and BAFsegmentation are comparable, while PSCN is fifty times slower than GFL (see Table 3.3).

In a clinical cancer study, it is often of great interest to determine whether a newly developed tumor is a recurrence (metastasis) of the original clone or an entirely new

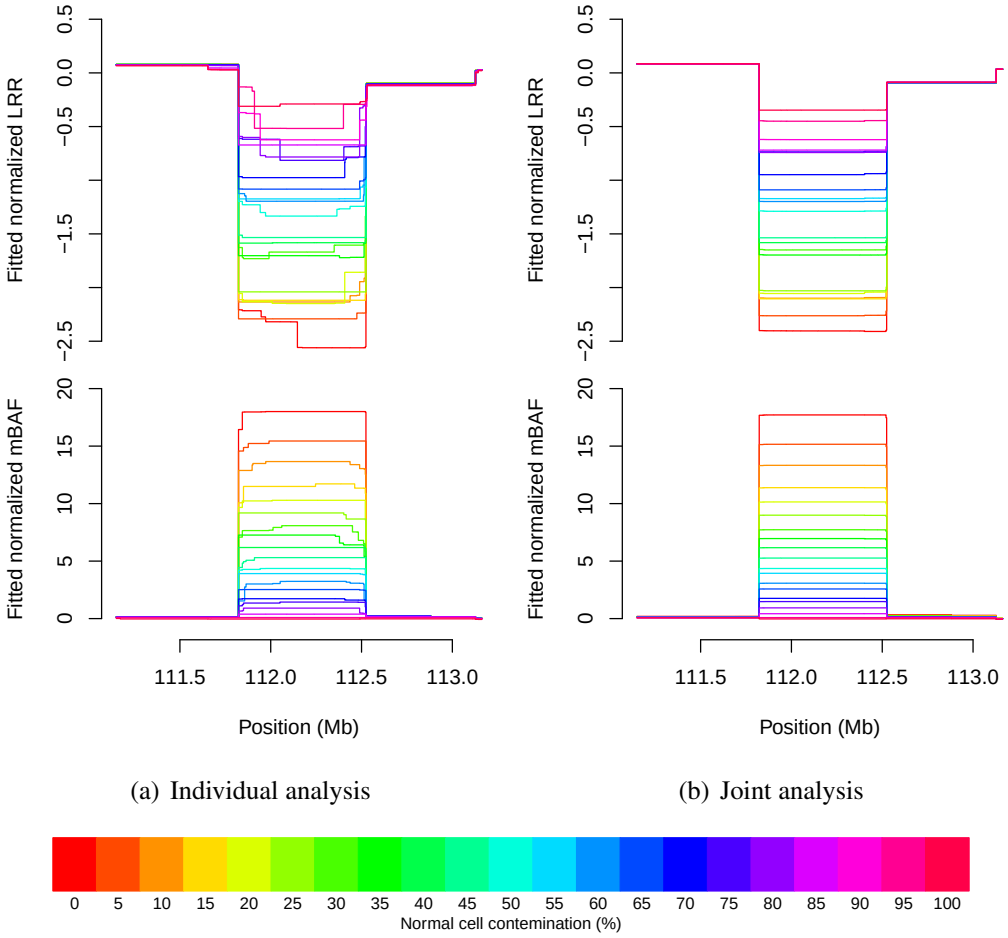
Table 3.3: Speed comparison of three methods: GFL, BAFsegmentation and PSCN.

Method	Time per sample in sec. (mean (std dev))
GFL	21.97 (1.31)
BAFsegmentation	41.73 (-)
PSCN	1154.18 (74.73)

type [Ostrovnaya et al. (2010)]. This application is further complicated by normal cell contamination in biopsy. The fitted patterns from our GFL model on the pooled data for a patient, collected from different positions at different times, may provide an intermediate and sparse interpretation of the original data, that can help to some extent in the decision. Taking this simulated tumor data set as an example, we can do a joint analysis on the combined data of all 21 samples. Both LRR and mBAF data are used. The joint analysis is done chromosome by chromosome. The tuning constants are $\lambda_1 = 0$, $\lambda_2 = 0.5 \times 3 \times \sqrt{\log N}$ and $\lambda_3 = 0.5 \times 3 \times \sqrt{42} \times \sqrt{\log N}$ for a chromosome N-SNP long. Figure 3.3 shows fitted models for a one-copy loss region on Chromosome 5q22, in comparison of individual analysis with different normal cell contamination levels and the joint analysis of all 21 samples. Joint analysis shows clearer patterns of consensus segmentations and boundaries, and interestingly, the spectrum of fitted segments reflect the gradient change of normal cell contamination levels (see Figure 3.3 (b)).

3.3.3 One sample assayed with multiple replicates and multiple platforms

We use the data from a study [Pinto et al. (2011)] assessing the performance of different array platforms and CNV calling methods to illustrate the advantages of joint analysis of multiple measurements on the same subject. DNA from four individuals was analyzed in triplicate on each of 5 platforms: Affymetrix 6.0, Illumina 1M, 660W, Omni1-Quad (O1Q) and Omni2.5-Quad (O2Q) (among others [Pinto et al. (2011)]). We use the results on the first three to define "true" copy numbers and try to reconstruct them using data from O1Q and O2Q. The nine "reference" experiments were analyzed with 4 or 5 CNV calling algorithms (see [Pinto et al. (2011)]) and a CNV was identified using majority votes: consistent evidence was required from at least 2 analysis tools, on at least 2 platforms, and in at least 2 replicates. Here CNVs detected in two replicates/algorithms/platforms are regarded as the same CNV and collapse down to one CNV with the outmost boundaries when they overlap with each other. Table 3.4 sum-



(a) Individual analysis (b) Joint analysis

Figure 3.3: Comparison of fitted profiles between analysis for each tumor sample with different normal cell contamination levels and joint analysis for all 21 tumor samples. Shown is a hemizygous loss on Chromosome 5q22. In each of the subplots, the upper panel shows the fitted profiles on LRR for each sample distinctly marked by a spectrum of colors, while the lower panel shows their corresponding fitted profiles on mBAF. Shown are data for heterozygous makers. (a) Individual analysis; (b) Joint analysis.

Table 3.4: Sample information and reference CNV regions summarized for each sample by their types and sizes. The ancestry of NA15510 was not recorded but inferred in [Korbel et al. (2007)]. PDR: Polymorphism Discovery Resource.

Sample	Gender	Ancestry	Resource	Type	$<$10k	10$-$50k	50$-$100k	$>$100k	Total
NA15510	Female	European	PDR	loss	12	25	3	7	47
				gain	0	0	1	4	5
				total	12	25	4	11	52
NA18517	Female	YRI	HapMap	loss	10	22	4	4	40
				gain	1	3	1	8	13
				total	11	25	5	12	53
NA18576	Female	CHB	HapMap	loss	13	16	4	5	38
				gain	0	2	2	4	8
				total	13	18	6	9	46
NA18980	Female	JPT	HapMap	loss	8	16	1	4	29
				gain	0	0	1	3	4
				total	8	16	2	7	33

marizes the information for each sample and the complied reference CNVs categorized by their type and size.

The test experiments are based on 1,020,596 and 2,390,395 SNPs on autosomes after some quality control, at a total of 2,657,077 unique loci. Since our focus here is to investigate how to best analyze multiple signals on the same subject, rather than on the specific properties of any CNV calling method, we carry out all the analyses using different settings of GFL in segmentation while keeping the same CNV calling and summarizing procedure. All segmentation is done on LRR only while calling procedure uses both LRR and BAF (with cut-off $r_1 = 10$ and $r_2 = 1$). Here we compare three segmentation settings to analyze these 6 experiments per subject (see Table 3.9 in Section 3.5.4 for more details about tuning parameters):

1. The signals from the three technical replicates with one platform are averaged

Table 3.5: Number of CNVs detected (Det.) and overlapping (Ovlp.) with reference results as well as average computation time for four samples under different analyses.

	NA15510		NA18517		NA18576		NA18980		
Analysis	# Det.	# Ovlp.	# Det.	# Ovlp.	# Det.	# Ovlp	# Det.	# Ovlp	Time (min.)
Analysis 1	170	38	144	34	160	25	145	22	1.2
Analysis 2	102	36	109	33	93	25	91	20	3.7
Analysis 3	80	38	82	32	69	25	56	15	8.5
MPCBS	98	34	88	28	59	18	68	21	313.9

and then segmented and subject to calling procedure separately. The final CNV list is the union of CNV calls from the two platforms.

2. The signals from the three technical replicates with one platform are each segmented and subject to calling procedure separately. A majority vote is used to summarize CNV result for each platform: a CNV needs to be called in at least two replicates out of three. The final CNV list is the union of the two platforms' results.

3. The signals from the three technical replicates of both platforms (6 LRR sequences) are segmented jointly. Calling procedure is still done on each replicate separately, and the same majority vote is used to summarize CNV result for each platform. Again, the final CNV list is the union of the two platforms' results.

To benchmark the result of joint analysis we use MPCBS [Zhang et al. (2010a)], a segmentation method, specifically designed for multi-platform CNV analysis. The segments output from MPCBS are proceeded to the same calling, majority voting, and summarizing procedure.

Table 3.5 presents the results: averaging results from different technical replicates leads to loss of power, while joint analysis of all the signals leads to the most effective performance. GFL joint analysis leads to results comparable to those of MPCBS, but it

is at least 30 times faster than the competing method.

3.3.4 Multiple related samples assayed with the same platform

In the context of a study of the genetic basis of bipolar disorder, the Illumina Omni2.5-Quad chip was used to genotype 455 individuals from 11 Columbian and 13 Costa Rican pedigrees. We use this data set to explore the advantages of a joint segmentation of related individuals. In absence of a reference evaluation of CNV status in these samples, we rely on two indirect methods to assess the quality of the predicted CNVs. We used the collection of CNVs observed in HapMap Phase III [Jakobsson et al. (2008)] to compile a list of 426 copy number polymorphisms (selecting all those CNVs with frequency ≥ 0.05 in pooled samples from 11 populations) and assumed that if we identify in our sample a CNV corresponding to one of these regions, we should consider it a true positive. For the purposes of this analysis we considered a detected CNV to correspond to one identified in HapMap if there was any overlap between the two regions.

Another indirect measure of the quality of CNV calls derives from the amount of Mendelian errors encountered in the pedigrees when we consider the CNV as a segregating site. De novo CNVs are certainly a possibility, and in their case Mendelian errors are to be expected. However, when the CNV in question is a common one (already identified in HapMap), it is reasonable to expect that it segregate in the pedigrees as any regular polymorphism. We selected a very common deletion on Chromosome 8 (HapMap reports overall frequency > 0.4 in 11 populations) and compared different CNV calling procedures on the basis of how many Mendelian errors they generate.

As mentioned before, PennCNV represents a state-of-the-art HMM method for the analysis of normal samples and, therefore, we included it in our comparisons. However, the parameters of the underlying HMM algorithm had not been tuned on the Omni2.5-Quad at the time of writing, resulting in sub-standard performance. Segmentation methods are less dependent on parameter optimization; hence, GFL analysis of

LRR and BAF one subject at a time can provide a better indication of the potential of single-sample methods. We considered two multiple-sample algorithms: GFL and MSSCAN [Zhang et al. (2010b)], both applied on LRR with group defined by pedigree memberships. (While a trio-mode is available in PennCNV [Wang et al. (2008)], this does not adapt to the structure of our families.) A final qualification is in order. While the authors of MSSCAN kindly shared with us a beta-version of their software, we find it not to be robust. Indeed, we were unable to use it to segment the entire genome. However, we successfully used it to segment Chromosome 8, so that we could include MSSCAN in the comparison based on Mendelian error rates.

Prior to analysis, the data was normalized using the GC-content correction implemented in PennCNV [Diskin et al. (2008)]. For individual analysis, the GFL parameters were $\lambda_1 = 0.1$, $\lambda_2 = 0$, and $\lambda_3 = 2 \times \sqrt{\log N}$, where N is the number of SNPs deployed on each chromosome; for pedigree analysis, the GFL parameters were $\lambda_1 = 0.1$, $\lambda_2 = 0.5 \times 2 \times \sqrt{\log N}$, and $\lambda_3 = 0.5 \times 2 \times \sqrt{0.3M} \times \sqrt{\log N}$, where M is the number of individuals in each pedigree. For MSSCAN, CNV size is constraint to be less than 200 SNPs and the maximum number of change points is set as 50. The calling procedure with $r_1 = 10$ and $r_2 = 1$ was applied to both the GFL and MSSCAN results.

Table 3.6 summarized the total number of copy number polymorphisms (CNPs) identified in our sample by different approaches and their overlap with known CNPs from HapMap. For the purpose of this comparison we considered as a CNP a variant with frequency at least 10% in our sample. All analysis modes of GFL agree more with HapMap list than PennCNV in the sense of percentage of overlap. It is also clear that GFL-pedigree analysis achieves larger overlap with HapMap data than GFL-individual analysis. The time cost per sample for pedigree is reasonable and scales well with the increment of sample size.

Table 3.7 summarizes the results of our investigation of a 154kb CNP region on Chromosome 8p (from 39,351,896 to 39,506,122 on NCBI Build 36 coordinate). All methods but PennCNV show detected deletions only; this coincides with the observa-

tion from HapMap data. We used option *Mistyping* of Mendel (version 11.0) [Lange et al. (2001); Sobel et al. (2002)] to detect Mendelian errors. Joint segmentation methods discover more hemizygous deletions than individual analysis, resulting in fewer Mendelian errors. MSSCAN discovers the largest number of hemizygous deletions. Figure 3.4 shows an example of large pedigree, where 3 out of 4 Mendelian errors are removed by joint analysis.

Table 3.6: The number of detected CNP regions with frequency ≥ 0.1 in our sample by different methods and their overlap with a list of CNP regions compiled from HapMap data. Computation time (in minute) is per sample.

Method	# Detected	# Overlap	% Overlap	Time (min.)
PennCNV	189	63	33.33%	3.44
GFL-Individual (LRR+BAF)	95	50	52.63%	3.90
GFL-Pedigree (LRR)	106	62	58.49%	1.57

Table 3.7: Detected copy numbers in a common deletion on Chromosome 8. Across the various algorithms, subjects are assigned to one of 4 types of copy number: for each algorithm, we report the total numbers of CN$\neq 2$ identified; the total number of "core" families with Mendelian errors; and the average computation time (in minute) per sample for the analysis of Chromosome 8.

Method	#CN=0	#CN=1	#CN=3	#families with errors	Time (min.)
PennCNV	125	39	102	35	0.19
GFL-Individual	123	97	0	20	0.21
GFL-Pedigree	123	137	0	15	0.09
MSSCAN-Pedigree	123	154	0	15	0.11

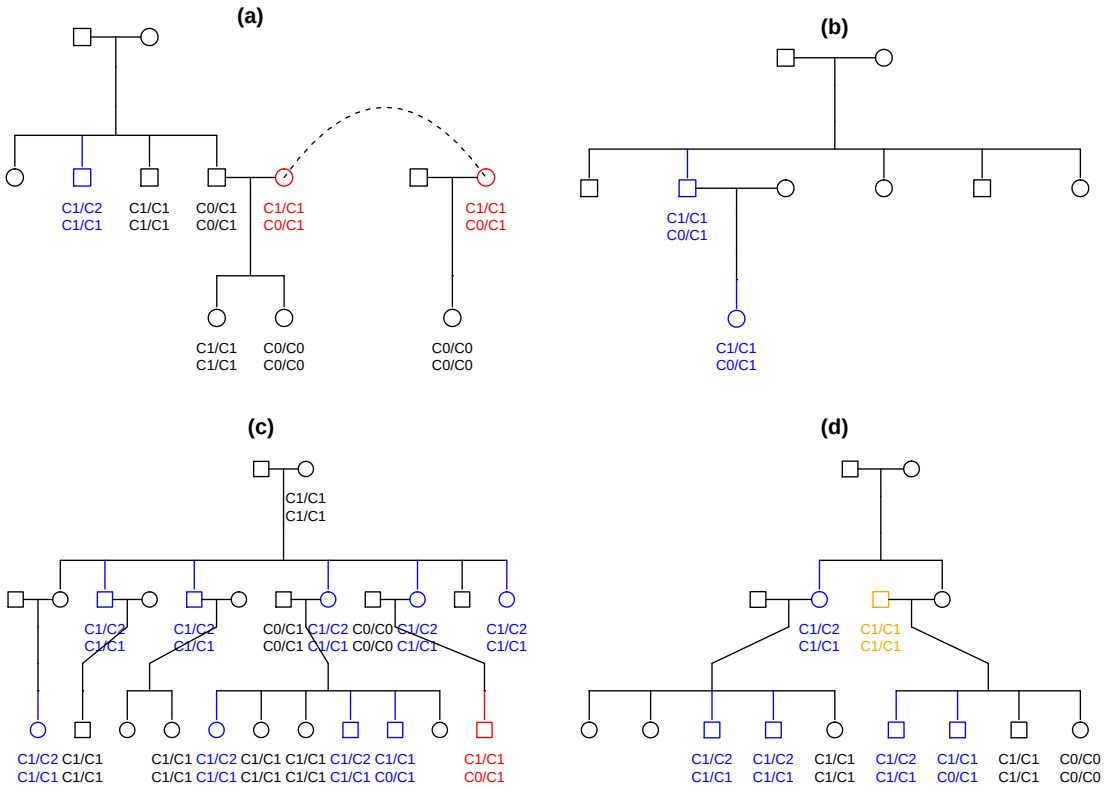


Figure 3.4: CNV detection and Mendelian errors for a Central American pedigree. Displayed are four extended families extracted from the big pedigree. Circles and squares correspond to females and males. Dashed line indicates the identical individual. Beneath each individual, from top to bottom, are CNV genotypes by PennCNV and by GFL. The subjects for whom PennCNV and GLF infer different CNV genotypes are highlighted in red and blue. Red indicates cases where the PennCNV genotype results in Mendelian error, while blue is for subjects where both genotypes are compatible with the rest of the family. Orange indicates a member for whom both PennCNV and GFL genotypes result in Mendelian error.

3.4 Conclusions

We have presented a segmentation method based on penalized estimation and capable of processing multiple signals jointly. We have shown how this leads to improvements in the analysis of normal samples (where segmentation can be applied to both total intensity and allelic proportion), tumor sample (where we are able to deal with contamination effectively), measurements from multiple platforms, and related individuals. Given that copy number detection is such an active area of research, it is impossible to compare one method to all the others available. However, for each of the situations we analyzed, we tried to select approaches that represented the most successful state-of-the-art. In comparison to these, the algorithm we presented performs well: its accuracy is always comparable to that of the most effective competitor and its computation time often more contained. We believe that for its versatility and speed, GFL is particularly useful for initial screening.

There are of course many aspects of CNV detection that we have not analyzed in this chapter: from normalization and signal transformation to FDR control of detected CNVs. There are also a number of improvements to our approach that appear promising, but at this stage are left for further work: for example, it is easy to modify algorithms so that the penalization parameters are location dependent to incorporate prior information on known copy number polymorphisms; more challenging is developing theory and method to select the values of these regularization parameters in a data-adaptive fashion.

Finally, while our scientific motivation has been the study of copy number variations, the joint segmentation algorithm we present is not restricted to specific characteristics of these data types, and we expect it will be applied in other contexts.

3.5 Appendix

Some technical details about justification of choice of tuning parameters in GFL and calling procedure as well as other supplementary information are listed as follows.

3.5.1 Bias estimation

Let x_{ij} be the data for sequence i at locus j after σ_i of each sequence is normalized to 1. With such normalization, the model (3.3) is reduced to a simpler form with global tuning parameters to each sequence for easier interpretation:

$$\begin{aligned} f(\boldsymbol{\beta}) &= \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N}(x_{ij}-\beta_{ij})^2 + \lambda_1\sum_{i=1}^{M}\sum_{j=1}^{N}|\beta_{ij}| \\ &\quad +\lambda_2\sum_{i=1}^{M}\sum_{j=2}^{N}|\beta_{ij}-\beta_{i,j-1}| + \lambda_3\sum_{j=2}^{N}\left[\sum_{i=1}^{M}(\beta_{ij}-\beta_{i,j-1})^2\right]^{\frac{1}{2}}. \end{aligned} \tag{3.9}$$

The solution to minimize $f(\boldsymbol{\beta})$ is unique for $f(\boldsymbol{\beta})$ is strictly convex. Denote the solution as $\hat{\boldsymbol{\beta}} = (\hat{\beta}_{ij})_{M\times N}$. Suppose sequence i is partitioned into $\hat{K}_i$ consecutive segments $\{\hat{R}_1^{(i)},\ldots,\hat{R}_{\hat{K}_i}^{(i)}\}$, delimited with change points $\hat{\mathcal{J}}_i = \{\hat{j}_1^{(i)},\ldots,\hat{j}_{\hat{K}_i-1}^{(i)}\} \subset \{2,\ldots,N\}$ (left end of segment $2,\ldots,\hat{K}_i$). The fitted means of each segment is denoted as $\hat{\boldsymbol{\mu}}^{(i)} = (\hat{\mu}_1^{(i)},\ldots,\hat{\mu}_{\hat{K}_i}^{(i)})$, i.e., $\hat{\beta}_{ij} = \hat{\mu}_k^{(i)}$, if $j \in \hat{R}_k^{(i)}$. The length (number of SNPs) of each segment is $\hat{L}_k^{(i)} = |\hat{R}_k^{(i)}|, k = 1,\ldots,\hat{K}_i$. Thus, the estimated mean vector for sequence i can be written as

$$\hat{\boldsymbol{\beta}}_i = \sum_{k=1}^{\hat{K}_i}\hat{\mu}_k^{(i)} I_{\hat{R}_k^{(i)}}.$$

$\hat{\boldsymbol{\beta}}$ is the optimal solution if and only if it satisfies the subgradient condition $\partial f(\hat{\boldsymbol{\beta}}) = 0$; that is,

$$\hat{\beta}_{ij} = y_{ij} - \lambda_1 s_{ij}^{(1)} - \lambda_2 s_{ij}^{(2)} - \lambda_3 s_{ij}^{(3)}, \tag{3.10}$$

where $s_{ij}^{(1)}$, $s_{ij}^{(2)}$ and $s_{ij}^{(3)}$ are coordinates of subgradient corresponding to β_{ij}'s appearing in each of the three penalty terms. Both bias estimation and asymptotic analysis rely on

the analytic form of subgradient. Now we discussed the bias induced by each penalty separately.

Bias induced by lasso penalty

It is easy to verify that the subgradient for the lasso penalty can be written as

$$s_{ij}^{(1)} = \text{sign}(\beta_{ij}),$$

where, with a bit abuse of notation,

$$\text{sign}(x) = \begin{cases} 1, & \text{if } x > 0, \\ -1, & \text{if } x < 0, \\ z \in [-1, 1], & \text{if } x = 0. \end{cases} \tag{3.11}$$

Hence, the lasso penalty term merely plays as a soft-thresholding on the fitted values resulted from the model (3.9) with $\lambda_1 = 0$, denoted as $\hat{\beta}_{ij}(0, \lambda_2, \lambda_3)$; that is, for any $\lambda_1 > 0$,

$$\hat{\beta}_{ij}(\lambda_1, \lambda_2, \lambda_3) = \text{sign}\left[\hat{\beta}_{ij}(0, \lambda_2, \lambda_3)\right] \left[\hat{\beta}_{ij}(0, \lambda_2, \lambda_3) - \lambda_1\right]_+,$$

where $(x)_+ = \max\{x, 0\}$. This is also highlighted in Lamma A.1 of [Friedman et al. (2007)] for model (3.9) with $\lambda_3 = 0$.

Bias induced by fused-lasso penalty

In model (3.9) with $\lambda_1 = 0$ and $\lambda_3 = 0$ (only fused-lasso penalty involved), Lemma 2.1 in [Rinaldo (2009)] gives an insightful characterization of $\hat{\boldsymbol{\mu}}^{(i)}$:

$$\hat{\mu}_k^{(i)} = \frac{1}{\hat{L}_k^{(i)}} \sum_{j \in \hat{R}_k^{(i)}} x_{ij} + \hat{c}_k^{(i)}, \quad k = 1, \ldots, \hat{K}_i,$$

where

$$\hat{c}_1^{(i)} = \begin{cases} -\frac{\lambda_2}{\hat{L}_1^{(i)}}, & \text{if } \hat{\mu}_2^{(i)} - \hat{\mu}_1^{(i)} > 0, \\ \frac{\lambda_2}{\hat{L}_1^{(i)}}, & \text{if } \hat{\mu}_2^{(i)} - \hat{\mu}_1^{(i)} < 0, \end{cases}$$

$$\hat{c}^{(i)}_{\hat{K}_i} = \begin{cases} \frac{\lambda_2}{\hat{L}^{(i)}_{\hat{K}_i}}, & \text{if } \hat{\mu}^{(i)}_{\hat{K}_i} - \hat{\mu}^{(i)}_{\hat{K}_i-1} > 0, \\ -\frac{\lambda_2}{\hat{L}^{(i)}_{\hat{K}_i}}, & \text{if } \hat{\mu}^{(i)}_{\hat{K}_i} - \hat{\mu}^{(i)}_{\hat{K}_i-1} < 0, \end{cases}$$

and, for $k = 2, \ldots, \hat{K}_i - 1$,

$$\hat{c}^{(i)}_{k} = \begin{cases} \frac{2\lambda_2}{\hat{L}^{(i)}_{k}}, & \text{if } \hat{\mu}^{(i)}_{k} - \hat{\mu}^{(i)}_{k-1} < 0, \hat{\mu}^{(i)}_{k+1} - \hat{\mu}^{(i)}_{k} > 0, \\ -\frac{2\lambda_2}{\hat{L}^{(i)}_{k}}, & \text{if } \hat{\mu}^{(i)}_{k} - \hat{\mu}^{(i)}_{k-1} > 0, \hat{\mu}^{(i)}_{k+1} - \hat{\mu}^{(i)}_{k} < 0, \\ 0, & \text{if } (\hat{\mu}^{(i)}_{k} - \hat{\mu}^{(i)}_{k-1})(\hat{\mu}^{(i)}_{k+1} - \hat{\mu}^{(i)}_{k}) > 0. \end{cases}$$

The result implies that the sample mean (as an unbiased estimate of true mean) of a local minimum/maximum segment (except it is located at either end) is shifted towards 0 due to fused-lasso penalty. The bias is positively proportional to λ_2 and negatively proportional to the length of the segment. It is more important to notice that there exists no configuration where a local minimum/maximum segment has a jump size (relative to neighboring segments) less than the amount of bias. It means that a CNV with small jump size or small length could possibly be merged into neighboring segments, if λ_2 is set too large.

Bias induced by group-fused-lasso penalty

The subgradient for group-fused-lasso penalty is given in the following Proposition 1.

Proposition 1: The β_{ij}'s involved in group-fused-lasso penalty have subgradient given by

$$s^{(3)}_{ij} = \begin{cases} -e_{i2}, & \text{if } j = 1, \\ e_{ij} - e_{i,j+1}, & \text{if } 1 < j < N, \\ e_{iN}, & \text{if } j = N, \end{cases} \tag{3.12}$$

for $i = 1, \ldots, M$, where $\mathbf{e}_j = (e_{1j}, \ldots, e_{Mj})^T$ for $j = 2, \ldots, M$ are given by

$$\mathbf{e}_j = \begin{cases} \left(\frac{\beta_{1j} - \beta_{1,j-1}}{||\boldsymbol{\beta}_{(j)} - \boldsymbol{\beta}_{(j-1)}||_{\ell_2}}, \ldots, \frac{\beta_{Mj} - \beta_{M,j-1}}{||\boldsymbol{\beta}_{(j)} - \boldsymbol{\beta}_{(j-1)}||_{\ell_2}} \right)^T, & \text{if } ||\boldsymbol{\beta}_{(j)} - \boldsymbol{\beta}_{(j-1)}||_{\ell_2} > 0, \\ \text{any } (e_{1j}, \ldots, e_{Mj})^T \text{ s.t. } ||\mathbf{e}_j||_{\ell_2} \leq 1, & \text{if } ||\boldsymbol{\beta}_{(j)} - \boldsymbol{\beta}_{(j-1)}||_{\ell_2} = 0. \end{cases} \tag{3.13}$$

Proof: The proof follows a similar technique used in the proof of Lamma A.1 in [Rinaldo (2009)]. Let $\mathbf{T} = [-\mathbf{I}_M, \mathbf{I}_M]$, where $\mathbf{I}_M$ is $M \times M$ identity matrix. Then, for any $2 \leq j \leq N$,

$$h(\boldsymbol{\beta}_{(j-1)}, \boldsymbol{\beta}_{(j)}) \triangleq ||\boldsymbol{\beta}_{(j)} - \boldsymbol{\beta}_{(j-1)}||_{\ell_2} = ||\mathbf{T}[\boldsymbol{\beta}_{(j-1)}^T, \boldsymbol{\beta}_{(j)}^T]^T||_{\ell_2}.$$

For the j such that $||\boldsymbol{\beta}_{(j)} - \boldsymbol{\beta}_{(j-1)}||_{\ell_2} > 0$, the sub-gradient is reduced to regular gradient, and thus can be derived in a usual way. We now focus on the j such that $||\boldsymbol{\beta}_{(j)} - \boldsymbol{\beta}_{(j-1)}||_{\ell_2} = 0$, i.e., the subgradient of β_{ij} at 0. By Cauchy-Schwartz inequality, we have

$$\begin{aligned} h(\boldsymbol{\beta}_{(j-1)}, \boldsymbol{\beta}_{(j)}) &\geq ||\mathbf{T}[\boldsymbol{\beta}_{(j-1)}^T, \boldsymbol{\beta}_{(j)}^T]^T||_{\ell_2} ||\mathbf{e}_j||_{\ell_2} \\ &\geq < \mathbf{T}[\boldsymbol{\beta}_{(j-1)}^T, \boldsymbol{\beta}_{(j)}^T]^T, \mathbf{e}_j > \\ &= h(\mathbf{0}) + < [\boldsymbol{\beta}_{(j-1)}^T, \boldsymbol{\beta}_{(j)}^T]^T - \mathbf{0}, \mathbf{T}^T \mathbf{e}_j > \end{aligned}$$

where $\mathbf{e}_j$ is any vector such that $||\mathbf{e}_j||_{\ell_2} \leq 1$. It follows by the definition of subgradient that $\mathbf{T}^T \mathbf{e}_j = [-\mathbf{e}_j^T, \mathbf{e}_j^T]^T$ is the subgradient for $[\boldsymbol{\beta}_{(j-1)}^T, \boldsymbol{\beta}_{(j)}^T]^T$. □

The bias induced by the group-fused-lasso penalty can be derived from the analytic form of subgradient accordingly and is given in the following Proposition 2.

Proposition 2: In model (3.9) with $\lambda_1 = 0$ and $\lambda_2 = 0$, the fitted means of segments for sequence i can be expressed as

$$\hat{\mu}_k^{(i)} = \frac{1}{\hat{L}_k} \sum_{j \in \hat{R}_k^{(i)}} x_{ij} + \hat{c}_k^{(i)}, \quad k = 1, \ldots, \hat{K}_i,$$

where

$$\hat{c}_k^{(i)} = \begin{cases} \frac{\lambda_3}{\hat{L}_1^{(i)}} \cdot r_i(\hat{j}_1^{(i)}), & \text{if } k = 1, \\ -\frac{\lambda_3}{\hat{L}_k^{(i)}} \cdot \left[r_i(\hat{j}_{k-1}^{(i)}) - r_i(\hat{j}_k^{(i)}) \right], & \text{if } 2 \leq k \leq \hat{K}_i - 1, \\ -\frac{\lambda_3}{\hat{L}_{\hat{K}_i}^{(i)}} \cdot r_i(\hat{j}_{\hat{K}_i - 1}^{(i)}), & \text{if } k = \hat{K}_i, \end{cases}$$

and

$$r_i(j) \triangleq \frac{\hat{\beta}_{ij} - \hat{\beta}_{i,j-1}}{||\hat{\boldsymbol{\beta}}_{(j)} - \hat{\boldsymbol{\beta}}_{(j-1)}||_{\ell_2}}.$$

Proof: The proof follows a similar technique used in the proof of Lemma 2.1 in [Rinaldo (2009)]. Following the subgradient condition (3.10) in case $\lambda_1 = 0$ and $\lambda_2 = 0$, we have

$$\hat{\mu}_k^{(i)} = \frac{1}{\hat{L}_k} \sum_{j \in \hat{R}_k^{(i)}} \hat{\beta}_{ij} = \frac{1}{\hat{L}_k} \sum_{j \in \hat{R}_k^{(i)}} x_{ij} - \frac{\lambda_3}{\hat{L}_k} \sum_{j \in \hat{R}_k^{(i)}} s_{ij}^{(3)}.$$

By Proposition 1 and simple algebra, we have

$$\sum_{j \in \hat{R}_k^{(i)}} s_{ij}^{(3)} = \begin{cases} -e_{i,\hat{j}_1^{(i)}}, & \text{if } k = 1, \\ e_{i,\hat{j}_{k-1}^{(i)}} - e_{i,\hat{j}_k^{(i)}}, & \text{if } 2 \le k \le \hat{K}_i - 1, \\ e_{i,\hat{j}_{\hat{K}_i-1}^{(i)}}, & \text{if } k = \hat{K}_i. \end{cases}$$

Note that at jump points, subgradient has explicit form as shown in Proposition 1. It follows that $e_{i,\hat{j}_k^{(i)}} = r_i(\hat{j}_k^{(i)})$, for $k = 1, \ldots, \hat{K}_i - 1$, where $r_i(\cdot)$ is defined in Proposition 2. □

Some interesting implications follow immediately. For sequence i, consider one of its fitted segment k with end points $[\hat{j}_{k-1}^{(i)}, \hat{j}_k^{(i)} - 1]$. If no other sequences share change points at these two ends, then the bias term $\hat{c}_k^{(i)}$ reduces to what it appears in model (3.9) with fused-lasso term only ($\lambda_1 = 0$ and $\lambda_3 = 0$). If m out of M sequences share change points at these two ends and also assume the jump size at these two locations for all the m sequences are roughly the same, then the absolute value of the bias term can be approximately written as $\frac{2\lambda_3}{\hat{L}_k^{(i)}} \cdot \frac{1}{\sqrt{m}}$. It means that if more than one sequences share change points at the same coordinate, then they can benefit from each other to reduce their individual bias, relative to the bias induced by fused-lasso penalty specific to each individual sequence.

3.5.2 Asymptotic behavior

Now we try to give a justification of the order of the magnitude of λ_2 and λ_3 in compatible with their large sample behavior, say, as $N \to \infty$. When the number of sequences M in segmentation task is relatively large, extra caution is needed for λ_3. Again, we

discuss asymptotic behavior of the solution influenced by fused-lasso and group-fused-lasso separately for easier exhibition.

Asymptotic behavior for fused-lasso penalty

In fused-lasso model ($\lambda_1 = 0$ and $\lambda_3 = 0$), the justification is directly inspired by the proof of Theorem 2.3 in [Rinaldo (2009)]. Denote the event

$$\mathcal{E}_i = \{\hat{\mathcal{J}}_i = \mathcal{J}_i\} \cap \{\text{sign}(\hat{\beta}_{ij} - \hat{\beta}_{i,j-1}) = \text{sign}(\beta_{ij} - \beta_{i,j-1}), \forall j \in \mathcal{J}_i\},$$

for $i = 1, \ldots, M$ respectively. This event means that all jump points and the direction of jumps are correctly identified for each sequence i. A necessary condition required for λ_2 is summarized in Proposition 3.

Proposition 3: It is required that $\lambda_2 = O(\sqrt{\log N})$ to ensure $\lim_{N\to\infty} \mathbb{P}(\mathcal{E}_i) = 1$ for $i = 1, \ldots, M$, at the linear rate.

This asymptotic behavior follows directly the proof of Theorem 2.3 in [Rinaldo (2009)]. We have some quick remarks:

1) If the signal of each sequence is not normalized, then $\lambda_{2,i} = c_2\sigma_i\sqrt{\log N}$, specific to sequence i.

2) In order to ascertain a CNV segment with length L and jump size δ, the bias needs to satisfy $\frac{2\lambda_{2,i}}{L} = \frac{2c_2\sigma_i\sqrt{\log N}}{L} < \delta$, i.e., $c_2 < \frac{1}{2\sqrt{\log N}} \cdot \frac{\delta}{\sigma_i}L$. Here, $\frac{\delta}{\sigma_i}$ can be interpreted as signal-to-noise ratio (SNR). For a specific platform, one may get a sense of the magnitude of SNR and L from prior knowledge. In practice, it is desired to take as large value of c_2 as possible to ensure the sparsity of the segmentation, but not too large in order to compensate for the constraint of signal strength ($\frac{\delta}{\sigma_i}L$). Based on our experiences of analysis of Illumina data, the results are not sensitive to the choice of c_2, provided that it falls into a reasonable range (see Section 2.2.6).

Asymptotic behavior for group-fused-lasso penalty

In group-fused-lasso model ($\lambda_1 = 0$ and $\lambda_2 = 0$), we have similar requirement of λ_3 as for λ_2, which is given in Proposition 4.

Proposition 4: It is required that $\lambda_3 = O(\sqrt{M}\sqrt{\log N})$ to ensure

$$\lim_{N\to\infty} \mathbb{P}(\cap_{i=1}^{M} \mathcal{E}_i) = 1,$$

at the linear rate.

Proof: For simplicity, we prove under the condition that ϵ_{ij} are i.i.d. $\mathcal{N}(0,1)$ (after σ_i is normalized to 1), while this condition can be relaxed [Rinaldo (2009)]. We also follow the same technique used in the proof of Theorem 2.3 in [Rinaldo (2009)]. Let $d_{ij} = \beta_{ij} - \beta_{i,j-1}$, $\hat{d}_{ij} = \hat{\beta}_{ij} - \hat{\beta}_{i,j-1}$, and $d^{\epsilon}_{ij} = \epsilon_{ij} - \epsilon_{i,j-1}$. Also denote $\mathbf{d}^{\epsilon}_j = (d^{\epsilon}_{1j}, \ldots, d^{\epsilon}_{Mj})^T$ and $\mathcal{J} = \cup_{i=1}^{M} \mathcal{J}_i$. By the subgradient condition (3.10), for each i, $\mathcal{E}_i$ holds if and only if

$$d^{\epsilon}_{ij} = \lambda_3(2e_{ij} - e_{i,j-1} - e_{i,j+1}), \quad \text{for } j \in \mathcal{J}_i^c, \tag{3.14}$$

and

$$|\hat{d}_{ij}| > 0, \quad \text{for } j \in \mathcal{J}_i. \tag{3.15}$$

Condition (3.15) has direct relevance to the bias issue, as discussed above. Now we focus on condition (3.14), which implies that

$$\max_{j\in\mathcal{J}^c} ||\mathbf{d}^{\epsilon}_j||_{\ell_2} = \max_{j\in\mathcal{J}^c} \lambda_3 ||2\mathbf{e}_j - \mathbf{e}_{j-1} - \mathbf{e}_{j+1}||_{\ell_2} < 4\lambda_3.$$

It is left to show that $\mathbb{P}(\max_{j\in\mathcal{J}^c} ||\mathbf{d}^{\epsilon}_j||_{\ell_2} \geq 4\lambda_3) = \mathbb{P}(\max_{j\in\mathcal{J}^c} ||\mathbf{d}^{\epsilon}_j/\sqrt{2}||^2_{\ell_2} \geq 8\lambda_3^2) \to 0$ as $N \to \infty$ for $i = 1, \ldots, M$. Note that for each j, $d^{\epsilon}_{1j}, \ldots, d_{Mj}$ are i.i.d. $\mathcal{N}(0,2)$, so

$||\mathbf{d}_j^\epsilon/\sqrt{2}||_{\ell_2}^2 \sim \chi_M^2$. Then we have

$$\begin{aligned}
& \mathbb{P}(\max_{j\in\mathcal{J}^c} ||\mathbf{d}_j^\epsilon/\sqrt{2}||_{\ell_2}^2 \geq 8\lambda_3^2) \\
= \; & \mathbb{P}(\cup_{j\in\mathcal{J}^c} ||\mathbf{d}_j^\epsilon/\sqrt{2}||_{\ell_2}^2 \geq 8\lambda_3^2) \\
\leq \; & \sum_{j\in\mathcal{J}^c} \mathbb{P}(||\mathbf{d}_j^\epsilon/\sqrt{2}||_{\ell_2}^2 \geq 8\lambda_3^2) \\
= \; & |\mathcal{J}^c| \mathbb{P}(||\mathbf{d}_j^\epsilon/\sqrt{2}||_{\ell_2}^2 \geq 8\lambda_3^2) \\
\leq \; & \exp\left[-\frac{1}{2}(8\lambda_3^2 - M) + \log|\mathcal{J}^c| - \frac{M}{2}\log\frac{M}{8\lambda_3^2}\right].
\end{aligned}$$

Here the first inequality is due to union bound and the second inequality is due to Chernoff's bound for χ_M^2 distribution. Under the assumption on sparsity of the change points, we have $|\mathcal{J}^c| = O(N)$ for fixed M. In our settings, M is fixed (which may rise up to thousands) while $N \to \infty$, yet in practice, M is not negligible with respect to $\sqrt{\log N}$. For example, $\sqrt{\log(10^6)} \approx 3.72$, and it is not uncommon to have more than 4 sequences for joint segmentation. Therefore, it is necessary to have $\lambda_3 = O(\sqrt{M}\sqrt{\log N})$. $\square$

We also have some remarks on how to determine λ_3:

1) If the signal of each sequence is not normalized, then $\lambda_{3,i} = c_3\sigma_i\sqrt{pM}\sqrt{\log N}$. The choice of p is decided case by case and discussed in the main text.

2) Following the above discussion about bias induced by group-fused-lasso penalty, if m out of M sequences carry CNVs with exactly the same boundary, the bias can be approximately written as $\frac{2c_3\sigma_i\sqrt{\log N}}{\hat{L}_k^{(i)}} \cdot \frac{\sqrt{pM}}{\sqrt{m}}$. On the one hand, if p is over estimated so that pM is much larger than m, the model would be over penalized and introduce more bias than that is attributed to individual fused-lasso penalty, and thus does not benefit from joint analysis; on the other hand, if pM is set too small, we have insufficient control on the sparsity of each sequence, so that it has to be compensated by the fused-lasso penalty. This is the reason why we need to incooperate $\rho(p)$ to re-weight the relative influence of the two penalties.

3.5.3 Details in calling procedure

We specify the likelihood functions of LRR and BAF signals in the log-likelihood ratio (3.8) as follows. For BAF signal, the likelihood is usually modeled for different copy number states as a mixture of densities surrounding a few possible BAF values corresponding to different genotypes [Colella et al. (2007); Wang et al. (2007)]. When population frequencies for allele A and B, p_A and p_B, are available or can be estimated from data, we have

$$L_{\text{BAF}}(x; c) = \sum_{s=0}^{c} \binom{c}{s} p_A^{c-s} p_B^s \phi_s(x; \mu_s, \sigma_s^2), \quad \text{for } c = 0, 1, 2, 3, 4,$$

where $\phi_s(\cdot; \mu_s, \sigma_s^2)$ is normal density for state s. The details in model and parameter specification are listed in Table 3.8.

Table 3.8: Model and parameter specification in BAF signal for each copy number state. $\hat{\sigma}_x$ is empirically estimated from BAF values in $(0.4, 0.6)$ for each individual.

c	s	Genotype	$\phi_s(\cdot)$	μ_s	σ_s
0	0	Null	normal	$1/2$	$10\hat{\sigma}_x$
1	0, 1	A, B	half normal	0, 1	$\hat{\sigma}_x$
2	0, 2	AA, BB	half normal	0, 1	$\hat{\sigma}_x$
	1	AB	normal	$1/2$	$\hat{\sigma}_x$
3	0, 3	AAA, BBB	half normal	0, 1	$\hat{\sigma}_x$
	1, 2	AAB, ABB	normal	$1/3$, $2/3$	$\hat{\sigma}_x$
4	0, 4	AAAA, BBBB	half normal	0, 1	$\hat{\sigma}_x$
	1, 2, 3	AAAB, AABB, ABBB	normal	$1/4$, $1/2$, $3/4$	$\hat{\sigma}_x$

In case where population frequencies p_A and p_B are not available, we might use an alternative likelihood function for BAF, defined by

$$L_{\text{BAF}}(x; c) = \max_{s \in \{0,\ldots,c\}} \phi_s(x; \mu_s, \sigma_s^2), \quad \text{for } c = 0, 1, 2, 3, 4,$$

where all parameters are defined in the same way (see Table 3.8).

For LRR signal, the likelihood function is simply defined by normal density:

$$L_{\text{LRR}}(y; c) = \phi(y; \mu_c, \sigma_c^2).$$

For $c = 0, 1, 3, 4$, μ_c and σ_c^2 are estimated based on the data $\mathbf{y}_R$ in segment R being considered, while μ_2 and σ_2^2 are estimated from the data of the whole chromosome on which segment R locates or, locally, from the data of a few hundred markers flanking the segment.

3.5.4 Additional results for multiple platform data

For the multiple platform data shown in Section 3.3.3, results from more CNV analyses and details about parameter settings in different analyses are supplemented in Table 3.9.

3.5.5 Software implementation

We have implemented the segmentation routine, which is our core contribution, in an R package (`Piet`) submitted to R-forge (http://r-forge.r-project.org). In Figure 3.5, we demonstrate a visualization of the CNV results on Chromosome 8 in the bipolar disorder study (see Section 3.3.4).

Table 3.9: Number of CNVs detected (Det.) and overlapping (Ovlp.) with reference results as well as average computation time for four samples under different analyses. Tuning parameters used in segmentation: $c_1 = 0.1$, $c_2 = 2$, $c_3 = 2$ and $p = 1$; ρ and M are specified for each analysis. Analysis A, C and E correspond to Analysis 1, 2 and 3 respectively in Table 3.5.

			NA15510		NA18517		NA18576		NA18980		
Analysis	ρ	M	# Det.	# Ovlp.	# Det.	# Ovlp.	# Det.	# Ovlp	# Det.	# Ovlp	Time (min.)
Analysis A: GFL done on averaged signal for each platform											
O1Q	1	1	92	34	73	22	71	21	69	20	0.3
O2Q	1	1	114	22	92	24	111	15	95	11	0.9
Union	-	-	170	38	144	34	160	25	145	22	1.2
Analysis B: GFL done on averaged signal of both platforms jointly											
	0	2	128	40	108	33	96	21	104	23	4.2
Analysis C: GFL done on three replicates separately for each platform											
O1Q	1	1	66	31	65	22	43	19	48	15	0.9
O2Q	1	1	68	23	65	22	65	12	59	13	2.8
Union	-	-	102	36	109	33	93	25	91	20	3.7
Analysis D: GFL done on three replicates jointly for each platform											
O1Q	0	3	64	32	66	22	54	21	53	18	1.1
O2Q	0	3	75	22	70	24	65	11	49	12	3.1
Union	-	-	106	36	115	33	96	22	83	21	4.2
Analysis E: GFL done on three replicates of both platforms jointly											
	0	6	80	38	82	32	69	25	56	15	8.5
MPCBS: Segmentation done on three replicates of both platforms jointly											
	-	-	98	34	88	28	59	18	68	21	313.9

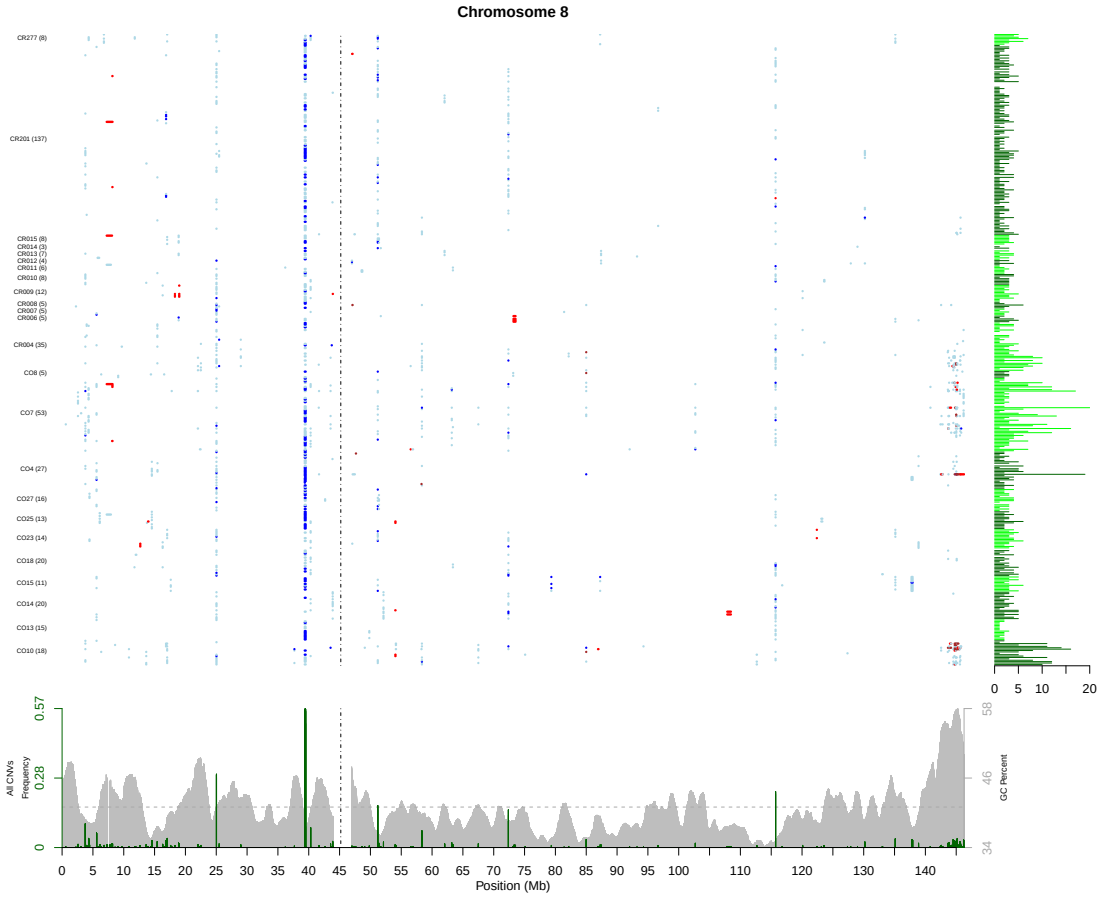


Figure 3.5: Visualization of pedigree-wise CNV analysis results of Chromosome 8 data in bipolar disorder study. In the main body of the plot, CNVs estimated for each individual are marked by small segments with color code: CN=0 in blue, CN=1 in light blue, CN=3 in red and CN=4 in brown. Each subject is a row, each SNP a column. Subjects belonging to the same pedigree are stacked together. The pedigree names are indicated on the left-hand side with the number of pedigree members included in parentheses. On the right-hand side, the barplot represents the number of CNV detected per subject. Two shades of green are switched alternately to indicate the pedigree to which the subject belongs. At the bottom, the gray histogram shows the GC content along the chromosome; coordinated with the representation of CNVs in the main body, the black histogram counts the frequency of CNV among the subjects represented. Vertical dotted line marks the centromere.

CHAPTER 4

Copy Number Polymorphisms in Two Central American Populations

In this chapter, we provide a detailed analysis of the genotype data obtained for 455 subjects from Costa Rica and Columbia. We begin with a brief description of the data set, followed by global and local admixture analysis, and a more comprehensive CNV investigation. Finally, we study characteristics of the detected CNPs in terms of their genomic locations and their frequencies in different populations.

4.1 Data description

This data set was generated in the context of a study with the main goal to identify genetic variation related to bipolar disorder (BP). The data contain 24 extended pedigrees, all with multiple individuals affected with severe BP (BP-I), and are collected from two related population isolates: the Central Valley of Costa Rica (hereafter CR) and the Antioquia Colombia (hereafter CO) [Carvajal-Carmona et al. (2003)]. Individuals from these pedigrees have been thoroughly phenotyped for a group of quantitative traits presumed to be related to BP as well as scanned for neurobehavioral observations. Currently, genotypes are available for 455 individuals assayed by the Illumina Omni2.5-Quad array with 2.45 million SNPs. Table 4.1 summarizes the data, illustrating how subjects are grouped into pedigrees and counting sample availability. For many of the analyses that follow, it is convenient to rely on a set of unrelated individuals. For this purpose, we have identified 67 subjects that are founders or married-ins in each

pedigree and treated as unrelated.

4.2 Admixture analysis

Population stratification has been well recognized as a confounder in genetic association studies. To remedy this, indications of ancestry for each subject are routinely included in an analysis. Subjects from Central America have admixed ancestry and it is important to evaluate what proportion of their genomes comes from each of the ancestral populations to account for overall genetic similarity between individuals. Moreover, differential prevalence of some diseases in the ancestral population can motivate mapping strategies based on identification of genomic segments in affected individuals that are derived from the same ancestry [Risch (1992); Patterson et al. (2004)].

The CR and CO populations were founded 300 to 400 years ago and grew in relative isolation, resulting in extensive linkage disequilibrium [Service et al. (2006)]. However, the genetic material that contributed to these populations is fairly close to that of present-day European, Native American and African populations, from which they separated only recently. Previous studies have indicated that the CR and CO populations are roughly 65-70% European, 20-25% Native American and 5-10% African in origin. The 2.45 million SNP array provides a highly dense coverage of the genome for admixture analysis in the BP study.

The admixture analysis can be performed in two layers, globally [Pritchard et al. (2000); Tang et al. (2005); Alexander et al. (2009)] and locally [Falush et al. (2003); Patterson et al. (2004); Tang et al. (2006); Sankararaman et al. (2008a,b); Paşaniuc et al. (2009)]. In the global admixture analysis, the main goal is to estimate what proportion of the entire genome is derived from each of the ancestral populations. In the local admixture analysis, a genome is regarded as a mosaic of consecutive segments with different ancestral origins. We aim at identifying the segment boundaries and assigning ancestral population to each segment.

Table 4.1: Summary of Costa Rica and Columbia (CR/CO) data set. The number of members and those genotyped of each pedigree is listed. These numbers are further subdivided by gender, affected status and sub-populations.

Costa Rican (CR)			Columbian (CO)		
Pedigree	# Subjects	# Genotyped	Pedigree	# Subjects	# Genotyped
CR1	77	35	CO1	35	18
CR2	36	5	CO2	19	15
CR3	11	5	CO3	25	20
CR4	27	5	CO4	13	11
CR5	23	12	CO5	30	20
CR6	30	8	CO6	19	14
CR7	13	6	CO7	15	13
CR8	15	4	CO8	37	16
CR9	39	7	CO9	53	27
CR10	26	3	CO10	89	53
CR11	17	8	CO11	13	5
CR12	230	137			
CR13	24	8			
Male	271	107	Male	155	86
Female	297	136	Female	193	126
Affected	64	49	Affected	84	72
Unaffected	504	194	Unaffected	264	140
CR Total	568	243	CO Total	348	212

4.2.1 Global admixture analysis

We employed ADMIXTURE package (version 1.04) to perform the global admixture analysis [Alexander et al. (2009)]. ADMIXTURE is an unsupervised admixture reconstruction algorithm that models the genome of the study population as a mixture of a pre-specified number of ancestral populations, and, for every individual in the sample estimates the proportion of genome attributable to each of the ancestral populations. Reference samples as a surrogate of ancestral populations are not required, but it is helpful to incorporate them in the analysis for a clearer interpretation of the results. For this purpose, we used CEU (Utah residents with Northern and Western European ancestry from the CEPH collection) and YRI (Yoruba in Ibadan, Nigeria) samples from HapMap release 3 [The International HapMap 3 Consortium (2010)], representing the European and African populations respectively, as well as a collection of Chibchan-Paezan speakers from Costa Rica-Panama border and Columbia acquired by our collaborators, representing the Native American (NA) population. More details about the sample size, SNP array used for genotyping and the number of SNP makers on autosomes for reference populations and our study population can be found in Table 4.2.

As we know a priori the genetic background of the samples from CEU, YRI and NA, we can interpret the ancestral populations of our study samples reconstructed by ADMIXTURE on the basis of how these reference samples are classified. The model in ADMIXTURE assumes linkage equilibrium between the markers, so the set of 52869 SNPs we used was derived from the set of 287543 common ones shared between the samples (Table 4.2), by pruning this set (keeping SNPs with $R^2 < 0.1$ in a 50-SNP sliding window, advanced by 10 SNPs in each move). We conducted the analysis using both all the 455 samples currently genotyped in the study and the set of 67 "unrelated" individuals (see Section 4.1). The result from all-sample analysis is not distorted appreciably by pedigree structure, as compared to that from "unrelated" samples, because the sample size of three reference populations is large enough so that the differentiation

Table 4.2: The genotype data used for admixture analysis. CEU and YRI samples from HapMap release 3 and samples acquired by collaborators from Costa Rica/Panama and Columbia are used to represent the three ancestral populations: European, African and Native American. Listed are the number of subjects, the genotyping platform used and the number of SNPs deployed on autosomes in each subset. CR: Costa Rica; CO: Columbia; NA: Native American.

Data set	# Subjects	Platform	# SNPs
Study sample	455	Illumina Omni2.5-Quad	2390395
CR	243		
CO	212		
NA	65	Illumina Human660W-Quad	538540
Costa Rica/Panama	51		
Columbia	14		
HapMap 3	225	Illumina Human1M & Affymetrix SNP 6.0	1389511
CEU	112		
YRI	113		
Merged data set	745		287543

between ancestral components dominates the finer differentiation between pedigrees. Therefore, we can safely report the results from the analysis including all individuals.

Table 4.3 summarizes the reconstructed genome-wide ancestry proportions for different subgroups. The estimated ancestral proportions confirm current knowledge on the populations under study. The CR population has a little more Native American component and less African and European components than the CO population. The ancestral proportions estimated for reference populations generally coincide with their nominal origins. However, it should be noted that the acquired Native American sample is not purely homogeneous – some individuals have a little portion of European ancestry. This may lead to larger variance in ancestry inference for our target populations. The estimated pattern of ancestral proportions for both reference populations and target populations can be visualized more clearly by a de Finetti plot (Figure 4.1). In Figure 4.1, the proximity of an individual to each vertex of the triangle indicates the proportion of the genome estimated to have ancestry in each of the three ancestral populations. The closer is one to the vertex, the more proportion does one have from corresponding ancestral population. Three reference populations concentrate on each of the three vertices, while the Native American samples spread out towards European vertex. While the CR and CO populations are generally close to European vertex, the CO individuals are more disperse towards African vertex.

To identify the genetic variants responsible for BP in CR and CO pedigrees, the possibility should be taken into account that the variants may be present with higher frequency in any of these three contemporary populations. The resolution provided by global admixture analysis is not adequate for this purpose. Therefore, it is necessary to evaluate locus-specific admixture proportions for all our subjects and use this information to specify the ancestral background of the variants.

Table 4.3: Estimated genome-wide ancestral proportions. The results are summarized with respect to different subgroups. Listed are percentages followed by standard deviations in parentheses. CR: Costa Rica; CO: Columbia; NA: Native American.

Data set		# Subjects	Native American	African	European
Study sample		455	23.3 (6.6)	3.8 (5.0)	72.9 (8.3)
CR		243	26.3 (4.2)	2.1 (2.3)	71.6 (4.7)
	Affected	49	26.2 (4.7)	2.2 (2.3)	71.6 (5.1)
	Unaffected	194	26.3 (4.1)	2.0 (2.4)	71.6 (4.6)
	Male	107	26.3 (4.5)	2.0 (2.3)	71.6 (4.9)
	Female	136	26.3 (4.0)	2.1 (2.3)	71.6 (4.6)
CO		212	19.9 (7.1)	5.7 (6.4)	74.4 (11.0)
	Affected	72	19.3 (6.9)	5.0 (6.0)	75.7 (10.5)
	Unaffected	140	20.3 (7.2)	6.1 (6.6)	73.7 (11.2)
	Male	86	19.5 (6.9)	5.5 (6.2)	75.0 (10.9)
	Female	126	20.2 (7.2)	5.8 (6.5)	74.0 (11.0)
NA		65	93.0 (11.2)	1.7 (3.6)	5.3 (8.6)
CEU		112	0.5 (0.4)	1.7 (0.9)	97.9 (0.9)
YRI		113	0.1 (0.2)	99.9 (0.4)	0.0 (0.4)

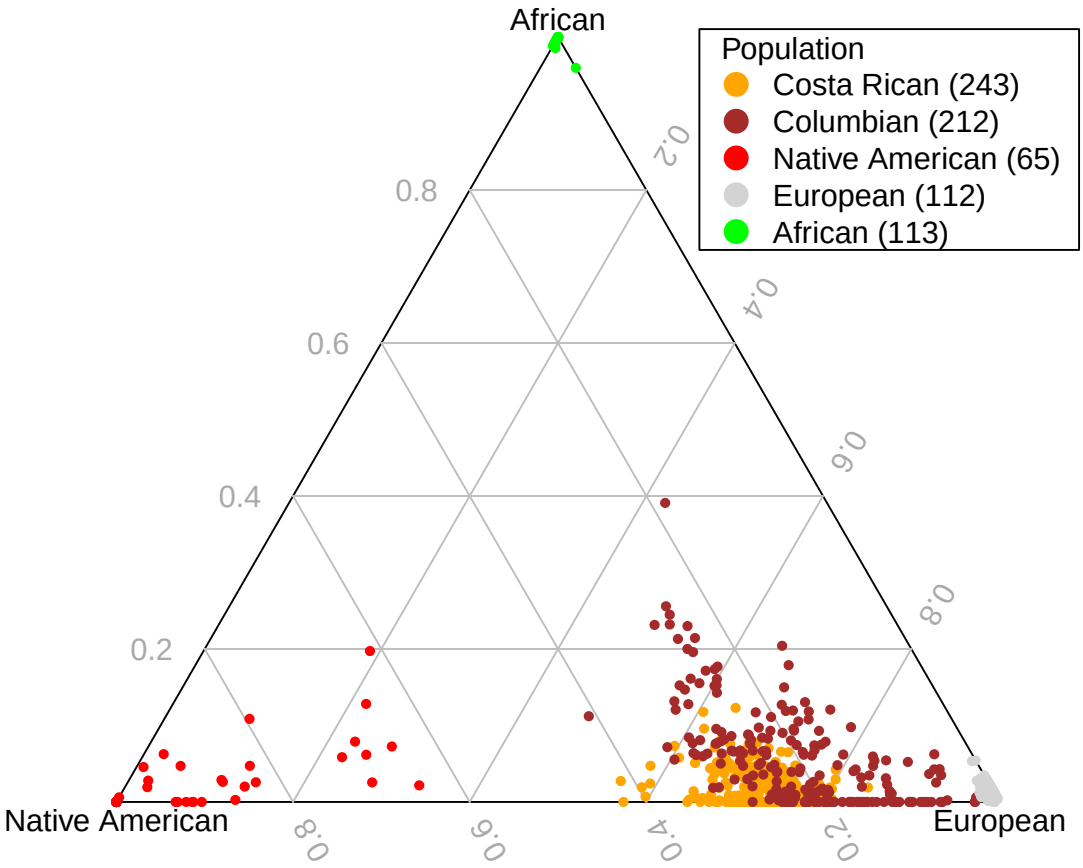


Figure 4.1: Visualization of global ancestry inference by de Finetti plot. Each dot represents an individual and colors are used to identify study samples. The sample size for each population is included in parentheses. Assuming three ancestral populations: Native American, European and African, the proximity of an individual to each vertex of the triangle indicates the proportion of the genome estimated to have ancestry in each of the three ancestral populations. Individuals from the same ancestral population are clearly clustering together at a common vertex.

4.2.2 Local admixture analysis

For local admixture analysis, we also used the data set as summaried in Table 4.2. We employed a package called LAMP (version 2.4 for 32 bit Linux), specifically designed for local ancestry inference [Sankararaman et al. (2008b); Paşaniuc et al. (2009)]. LAMP uses a sliding window strategy to divide the entire genome into small segments, on which a clustering algorithm is applied to assign ancestral labels to each individual. The SNPs within each window are assumed to have the same ancestral assignment. The window length should be short enough so that there are less likely to be breakpoints within the window and long enough so that there is enough information for accurate clustering. An optimal choice is related to the ancestral proportions, the number of generations since the beginning of mixing process, and the recombination rate and is determined a priori. Following the clustering step, a majority vote is applied for each SNP, over all windows that overlap with the SNP, in determination of the most likely ancestral assignment. This divide-and-conquer type of strategy dramatically reduces the complexity of the problem, while still keeping the accuracy of inference [Sankararaman et al. (2008b)].

Our analysis was conducted chromosome by chromosome. We only estimated the ancestral proportions of CR and CO individuals. The data from reference populations was instead used to estimate the minor allele frequency at each SNP for ancestral populations. The initial proportions of three ancestries (Native American, European and African) are taken as the average over the 67 "unrelated" samples in global admixture analysis (see Section 4.2.1), which are 0.25, 0.69 and 0.06 respectively. The number of generations along the mixing history is set as 20, an estimated upper bound for our samples as suggested by LAMP. The recombination rate is set to be 10^{-8} per base pair per generation as recommended. LAMP assumes SNPs within a window to be independent given ancestral labels. In practice, this is achieved by the greedy removal of SNPs having correlation coefficient $R^2 > 0.1$ from the window.

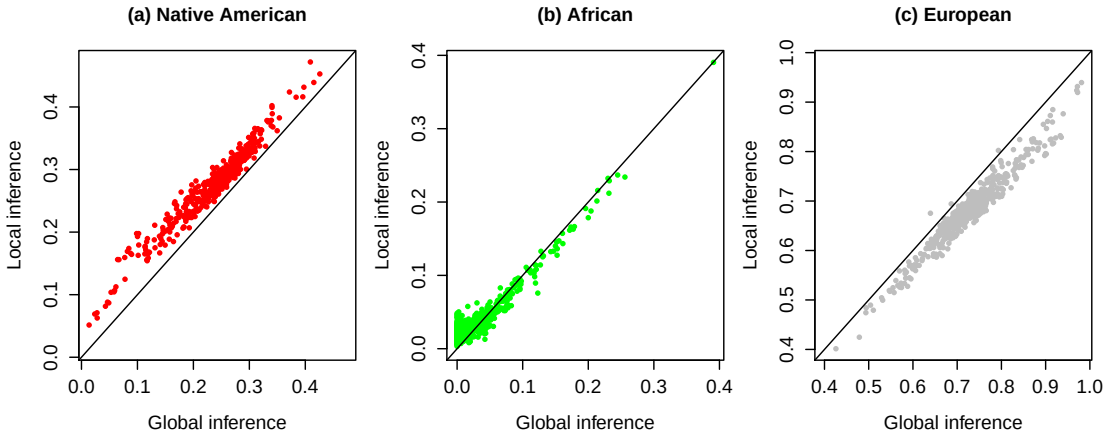


Figure 4.2: Comparison between global and local admixture analyses. Three ancestral components, Native American, European and African, are color-coded in red, green and gray. In each subplot, a dot represents an individual, whose ancestral proportion estimated by local inference is plotted against estimate from global inference. The local ancestral proportion is generated by averaging local estimation over the genome for each individual. The solid line indicates $y = x$.

The local ancestral inference for each individual is summarized by averaging it over the entire genome. The averaged ancestral proportions are reported for each subgroup in Table 4.4 in the same way as global admixture analysis (see Table 4.3). We also compared the results from global and local admixture analysis (see Figure 4.2). Overall, the results from local analysis are consistent with those from global estimate, whereas the local inference results in higher Native American component and lower European component.

To compare the genetic differences between the two populations, we summarized ancestry profiles for CR and CO samples separately, by averaging the estimated numbers of ancestral alleles for each SNP across the samples with respect to each of the three ancestral populations. In Figure 4.3, the ancestry profiles for CR and CO samples are plotted according the genomic locations of the SNP markers and compared for each of the three ancestral components. Overall, the CR and CO profiles are very similar to

Table 4.4: Estimated genomic ancestral proportions derived by local inference. The three ancestral proportions for each individual are generated by averaging local estimation over the genome. The results are then summarized with respect to different subgroups. Listed are percentages followed by standard deviations in parentheses. CR: Costa Rica; CO: Columbia.

Data set		# Subjects	Native American	African	European
Study sample		455	27.3 (6.0)	4.5 (4.2)	68.2 (7.5)
CR		243	29.4 (4.3)	3.2 (1.4)	67.4 (4.7)
	Affected	49	29.4 (5.1)	3.2 (1.2)	67.4 (5.3)
	Control	194	29.5 (4.1)	3.2 (1.4)	67.4 (4.5)
	Male	107	29.5 (4.6)	3.3 (1.4)	67.3 (5.0)
	Female	136	29.4 (4.1)	3.1 (1.4)	67.5 (4.5)
CO		212	24.9 (6.7)	6.0 (5.7)	69.2 (9.7)
	Affected	72	24.3 (6.4)	5.5 (5.4)	70.2 (9.3)
	Control	140	25.1 (6.9)	6.2 (5.9)	68.6 (10.0)
	Male	86	24.3 (6.5)	5.9 (5.4)	69.7 (9.8)
	Female	126	25.2 (6.8)	6.0 (5.9)	68.8 (9.7)

Table 4.5: Correlation coefficient between average ancestral profiles of two groups of individuals. 95% Confidence interval is included in parentheses. Two comparisons are listed: CR versus CO populations and affected individuals versus controls.

Ancestry	CR vs CO	Affected vs Control
Native American	0.1711 (0.1485, 0.1936)	0.7508 (0.7399, 0.7614)
African	0.4782 (0.4511, 0.5044)	0.8612 (0.8511, 0.8707)
European	0.2420 (0.2207, 0.2631)	0.7854 (0.7761, 0.7944)

each other, while some changes in ancestry are observable across the genome. These variations in average ancestry are probably attributable to inaccurate reconstruction due to limited information. We also compared average local ancestry proportion in between BP-affected and unaffected individuals. The difference in the average ancestral profiles are even smaller than the difference between CR and CO populations. These observations are also reflected in the correlation coefficient between the profiles of the two groups in comparison (see Table 4.5). The correlation between affected and unaffected individuals is much larger than the correlation between CR and CO populations, for each of the three ancestral components.

Locus-specific ancestry is more informative in that it is able to specify the genetic background to which variants of interest belong. We will use this information to evaluate the ancestral background of detected CNP regions in our study.

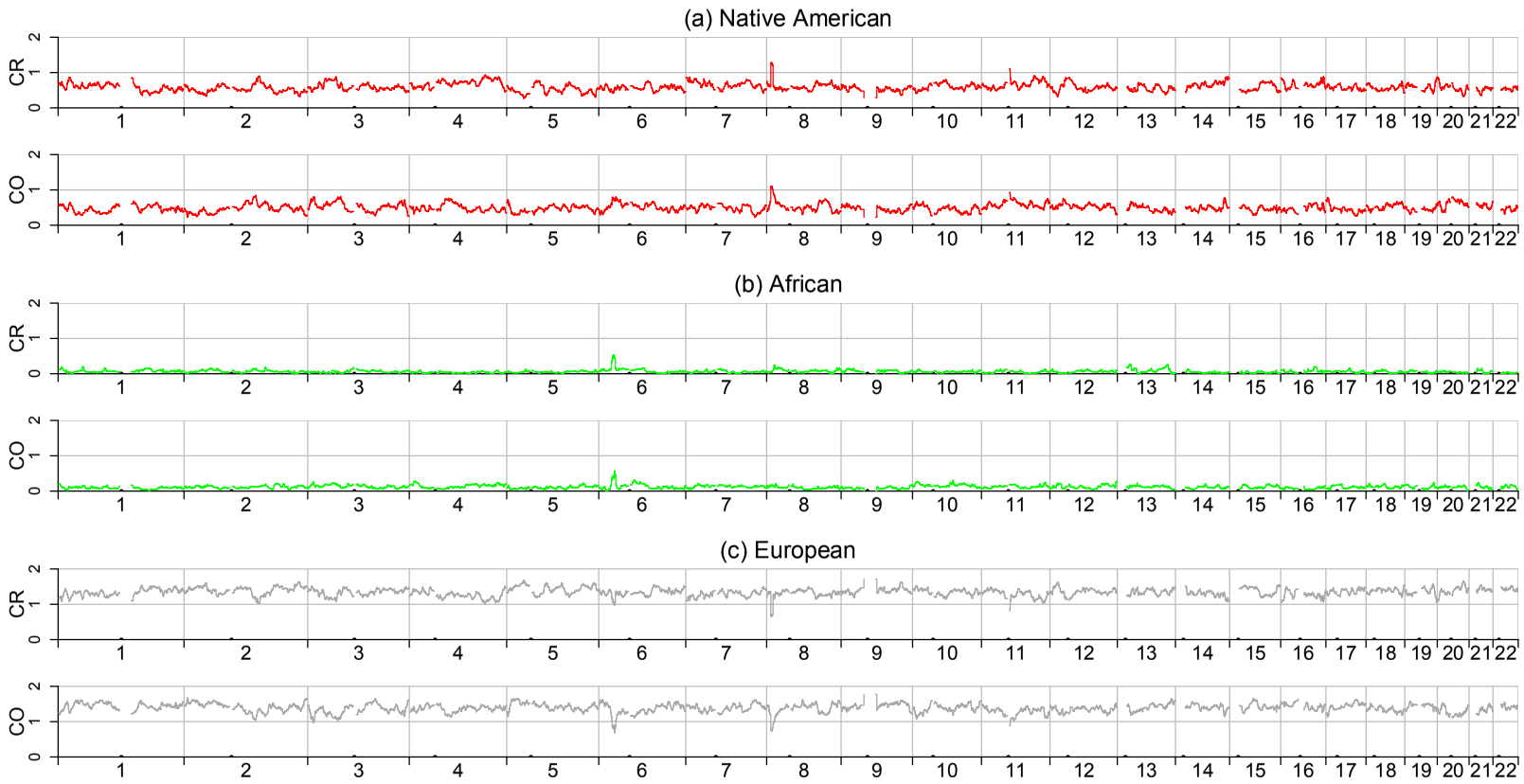


Figure 4.3: Average local ancestry profile for CR and CO samples. Three ancestral components, Native American, European and African, are color-coded in red, green and gray. In each subplot, the estimated number of ancestral alleles at each SNP location is averaged across CR and CO samples separately, and displayed against genomic positions. Vertical gray lines delimit chromosomes while small dots indicate the location of centromeres.

4.3 CNV analysis

4.3.1 CNV detection and quality control

We did CNV analysis based on the data of 2390395 SNPs on 22 autosomes for the 455 individuals as described in Section 4.1. Prior to CNV analysis, the total intensity signals (LRR) from all individuals were subject to an adjustment using singular value decomposition (SVD): the long-range spatial variation was captured by the first principal component and was then subtracted; the residuals were used in subsequent analysis [Zhang et al. (2010b); Siegmund et al. (2011)]. We used the segmentation method proposed in Chapter 3 for joint analysis on pedigrees. Only LRR information was used in segmentation step while both LRR and BAF signals were used in calling procedure (see Section 3.2.5). Segmentation was done chromosome by chromosome with the tuning parameters $\lambda_1 = 0.1$, $\lambda_2 = 0.5 \times 2 \times \sqrt{\log N}$, and $\lambda_3 = 0.5 \times 2 \times \sqrt{0.3M} \times \sqrt{\log N}$, where N is the number of SNPs deployed on each chromosome; M is the number of individuals in each pedigree.

This analysis resulted in a list of 20523 CNVs, which was then filtered by a set of criteria: a CNV must harbor at least 5 SNPs and be at most 1 Mb in size. The log-likelihood ratio (3.8) defined in calling procedure (see Section 3.2.5) was taken as an additional criterion (hereafter called confidence score and denoted by s). In Figure 4.4, CNVs are subgrouped by their copy numbers and for each CNV, its confidence score is plotted against the number of SNPs within the CNV region (hereafter denoted by t). We can roughly see a linear correlation between s and t. Also it is worth noting that the variance of s is approximately linear in t, due to the additive form of the log-likelihood ratio (3.8). For n CNVs belonging to a specific copy number group, we assume

$$s_i = \beta t_i + \epsilon_i, \quad i = 1, \ldots, n \tag{4.1}$$

where ϵ_i are i.i.d. $\mathcal{N}(0, t_i\sigma^2)$. The nominal point-wise 95% lower confidence bound $\hat{\beta}t_i - q_{0.95}\sqrt{t_i}\hat{\sigma}$ was then taken as a threshold to remove the CNVs lack of confidence.

Table 4.6: Summary statistics for detected CNVs.

Statistic	CN=0	CN=1	CN=3	CN=4	Overall
# CNVs	2043	14823	1731	519	19116
# CNVs per individual	4.5	32.6	3.8	1.1	42.0
Mean size (kb)	26.68	32.90	135.45	117.55	43.82
Median size (kb)	8.64	11.05	62.62	56.05	12.55

Here $q_{0.95}$ is the 95 percentile of $\mathcal{N}(0,1)$; $\hat{\beta}$ and $\hat{\sigma}$ are maximum likelihood estimates. Figure 4.4 shows the estimated 95% lower confidence bound (dashed line) for each copy number status. Those CNVs below the bound were dropped out of subsequent analysis.

This finally resulted in 19116 CNVs after quality control. Table 4.6 reports several summary statistics such as the number of detected CNVs, the number of CNVs per individual, mean and median size, which are subdivided by copy number status. The detected CNVs with CN=0 have the smallest mean/median size, followed by CN=1, CN=4 and CN=3 in an increasing order of mean/median size, reflecting indirectly the easiness with which a CNV with different copy numbers could be detected. The same trend is also observed in $\hat{\beta}$, estimated for each copy number status (see Figure 4.4), which can be considered as a surrogate measurement of CNV signal level carried by one SNP.

4.3.2 Copy number polymorphism

Construction of CNP regions

Summarizing the results of our CNV analysis of the study sample, we complied a list of 685 CNPs as follows: we stacked the detected CNVs to decide the boundaries of CNP regions, selected the ones that occur in more than 1% of the 455 in-

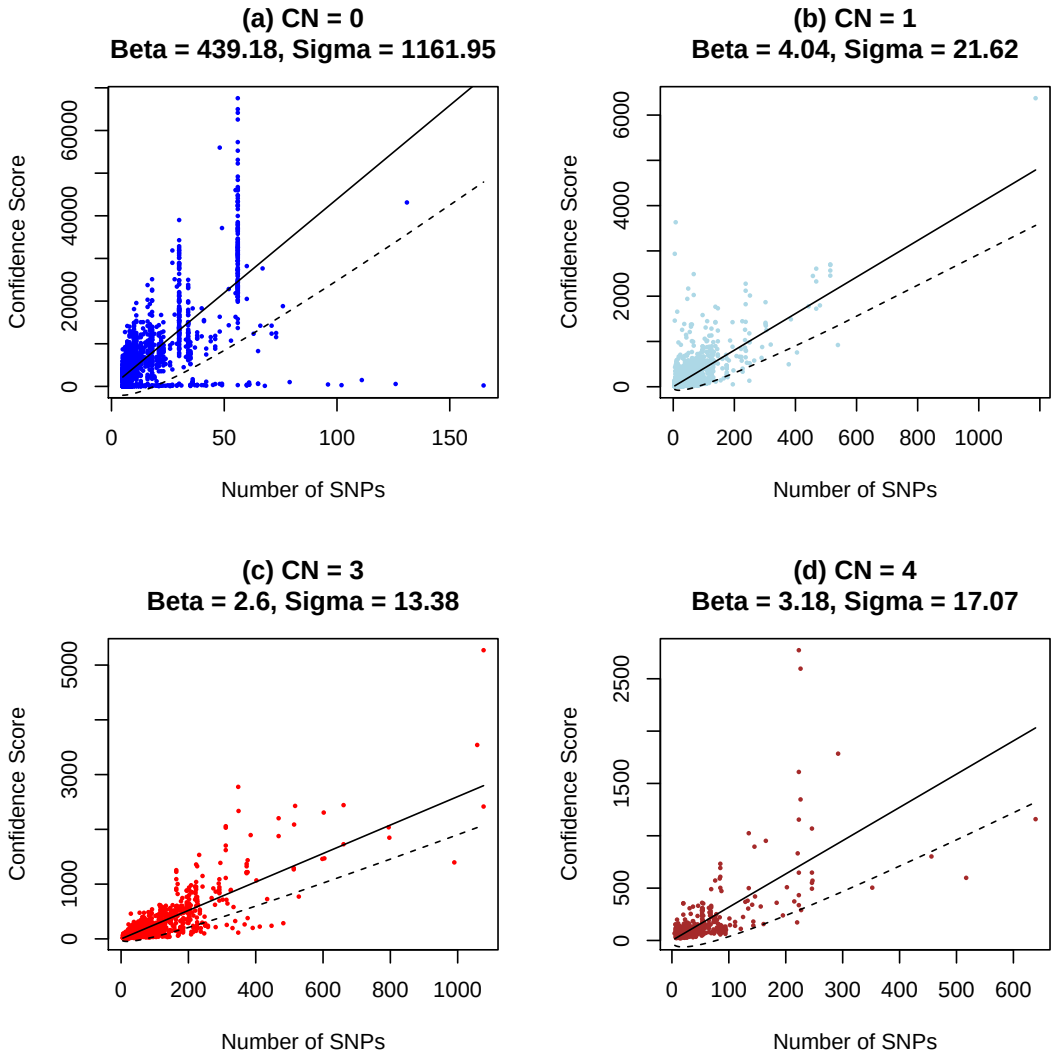


Figure 4.4: Quality control for detected CNVs. The dots with different colors correspond to CNVs of different copy numbers, subdivided in each plot: (a) CN=0 in blue; (b) CN=1 in light blue; (c) CN=3 in red; and (d) CN=4 in brown. Confidence score is plotted against the number of SNPs within CNV segment. Estimated parameters β and σ in equation (4.1) are given in the title of each subplot. Solid line indicates the fitted linear function, while dashed curve indicates the point-wise 95% lower confidence bound. Those dots below the dashed curve were filtered out from subsequent analysis.

dividuals and kept the ones that encompass at least 5 SNPs. The list was further compared with known CNPs retrieved from two resources: the database of genomic variants (DGV) – a curated catalogue of structural variation in healthy human samples [Zhang et al. (2006)], and the database of genomic structural variation (dbVar) – a database of structural variation which collects variant data from studies submitted for different organisms. More specifically, the data from DGV was downloaded at `http://projects.tcag.ca/variation/downloads/` on March 16, 2012. This version (10.Nov.2010) contains 66741 structure variants, including CNVs and inversions, complied from 42 studies. A total of 27644 common CNVs on autosomes remained after excluding all inversions and the CNVs, that have been observed only once or have frequency < 0.01. The remaining list was curated by reducing redundancy (taking the inner-most boundaries of CNVs overlapping with each other) and excluding extremely small CNVs (with size $<$ 1kb), resulting in a list of 10745 CNVs. The data from dbVar was downloaded at `http://www.ncbi.nlm.nih.gov/dbvar/` on March 16, 2012, under the accession number nstd46 [Campbell et al. (2011)] . It contains 1183 CNPs in human genome, among which 1111 are located on autosomes. After filtering this list by the similar criteria as above, we obtained a refined list of 995 CNP regions. All genomic coordinates used are based on NCBI Build 36 assembly. Among the 685 CNPs detected in CR/CO sample, 446 are documented in either DGV or dbVar. Note that this is a lower bound of the overlap due to the previous curation of the CNVs in these databases. If we use the originally documented CNVs, we will have an overlap of 530. Therefore, the set of 446 CNPs should be interpreted as a very stringent and conservative subset of the detected CNPs.

Using a simple estimate of the frequency for each CNP based on all 455 individuals treated as independent or the subset of 67 ”unrelated” individuals (see Section 4.1), one can note that the CNPs documented in previous studies generally have higher frequency than the remainder (see Figure 4.5). This is not surprising, as they were widely found in other populations. However, It is also interesting to look deeper into the other 239

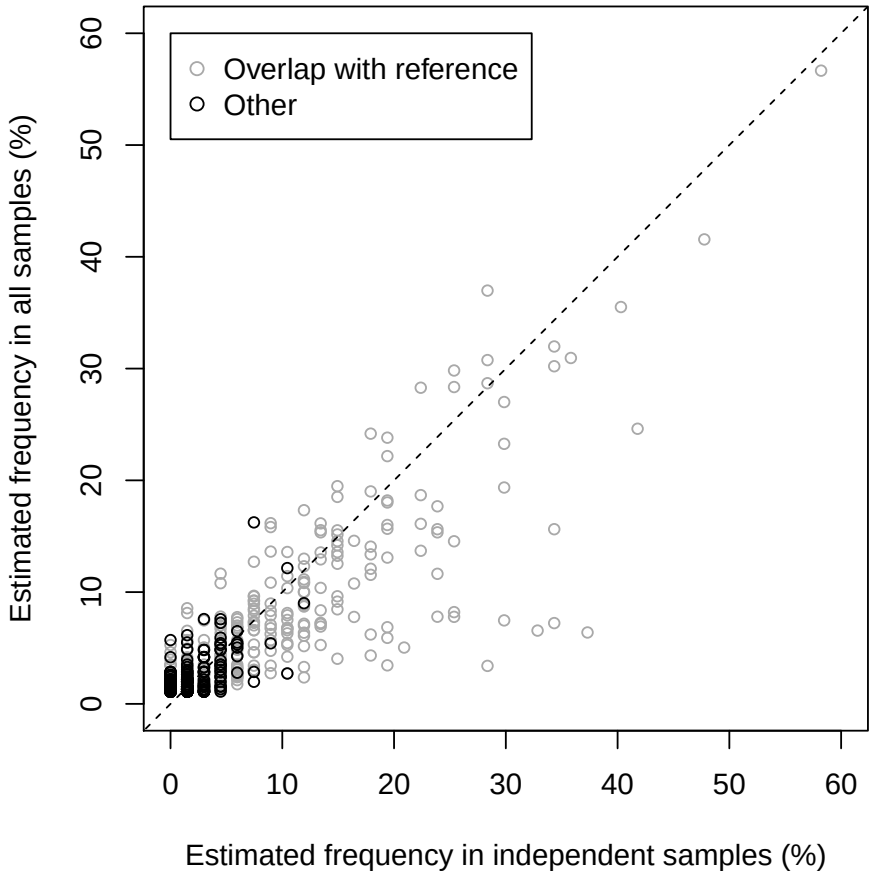


Figure 4.5: Frequency of CNPs detected in CR/CO sample. For each CNP, the rough estimation of frequency based on all 455 subjects is plotted against the estimation based on 67 "unrelated" individuals extracted from the CR/CO pedigrees. The 446 CNPs also documented in DGV or dbVar are displayed in gray while the other ones are in black.

CNPs, some of which may be specific to our study population, in future work. In the remainder of the chapter, unless specified otherwise, we focus on the analysis of the 446 CNPs, in an attempt to base subsequent inference mainly on well validated polymorphisms. Figure 4.6 illustrates the location and characteristics of these CNPs.

Check Mendelian errors

It is reasonable to consider CNPs, especially those that have been discovered in multiple populations, as segregating sites in pedigrees rather than de novo mutations. An

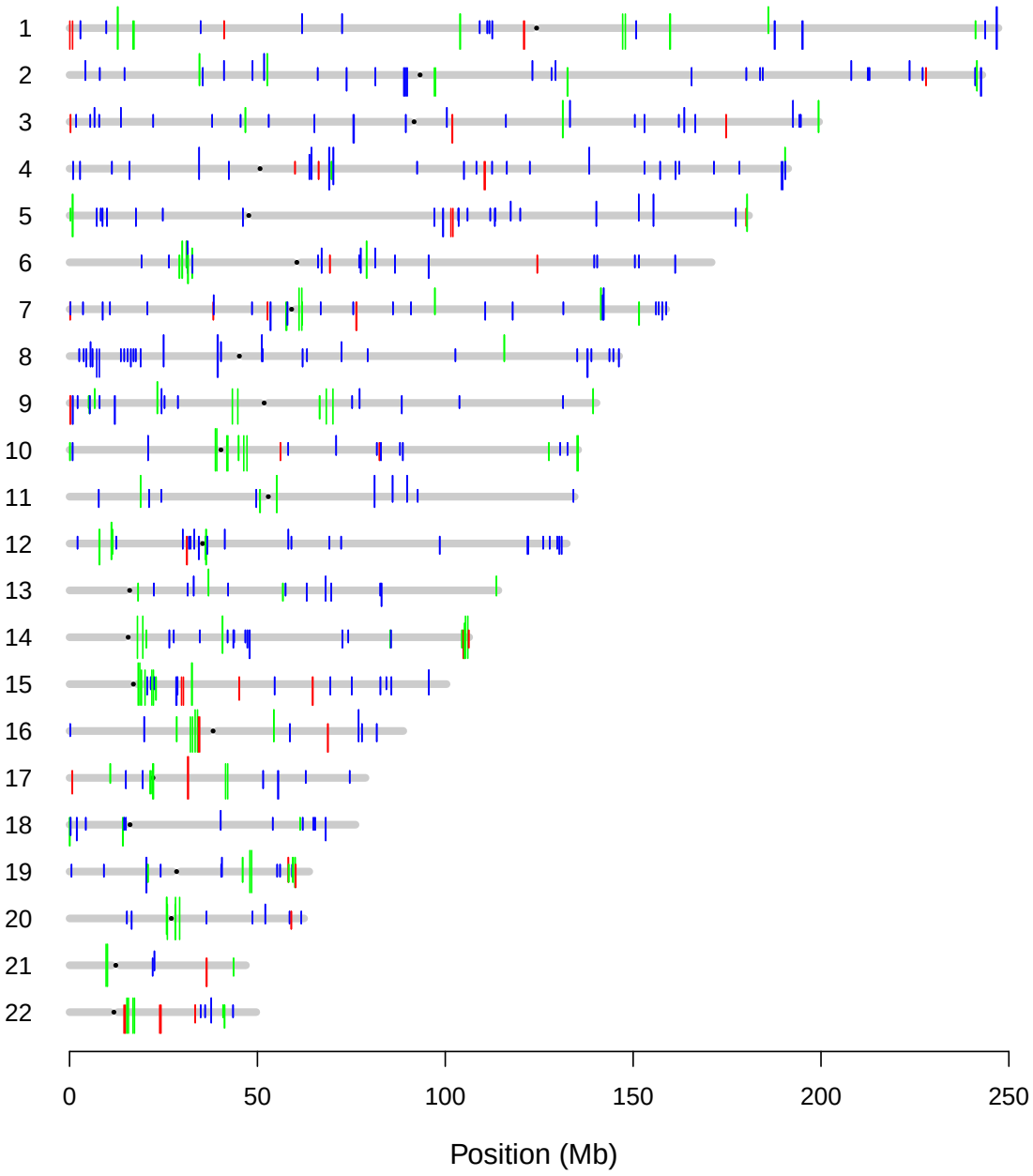


Figure 4.6: Visualization of detected CNPs. CNPs are plotted in short vertical segments according to their physical locations on each chromosome, which is manifested by thick gray line with black dot indicating the centromere. Deletion polymorphism is coded in blue, duplication in red and regions with both types of polymorphism in green. The length of the upper portion of each vertical segment above the gray line is positively related to the population frequency of corresponding CNP, whereas the length of the lower portion is positively related to the size of the CNP.

inconsistency in transmission is most likely attributed to Mendelian error, due to inaccurate detection of CNV. Therefore, the amount of Mendelian errors can be used as an indirect assessment of the quality of our CNV calls and a criterion to eliminate suspicious CNV calls from downstream analysis. We divided the pedigrees into nuclear families (parents and their offsprings), among which 162 have genotyped members, and accounted them as the basic unit in measuring Mendelian errors. We used the option *Mistyping* of Mendel (version 12.0) [Lange et al. (2001); Sobel et al. (2002)] to detect Mendelian errors for each family.

Figure 4.7 (a) shows that the number of families in which Mendelian error was detected for a CNP generally increases with the number of subjects estimated to carry the CNP. This is not surprising, as more CNP carriers provide more opportunities for Mendelian errors to occur. However, if we consider the measure of error rate defined by he ratio of the number of families with Mendelian errors among the number of families with CNP carriers, one notices that CNPs with larger number of carriers tend to have lower error rate. This is most likely the result of our joint analysis method, which has increased accuracy for more common CNPs.

Estimation of allele frequency of CNV genotype

To properly take into account the dependency among the observations due to pedigree structure, while taking advantage of our entire data set, we used the option *Allele frequencies* of Mendel (version 12.0) [Lange et al. (2001, 2005)] to estimate the allele frequencies of CNV genotypes for each CNP region. Here we consider five allelic types: C0, C1, C2, C3 and C4, indicating 0 to 4 copies of the CNP segment. A certain copy number may correspond to one or more genotypes. For example, a deletion with CN=0 could have only one possible genotype: C0/C0, while a duplication with CN=4 could have three possible genotypes: C0/C4, C1/C3, C2/C2. The Mendel option for allele frequency estimation has two models: one takes pedigree structure into account and respects relationships between members, while the other simply treats all pedigree

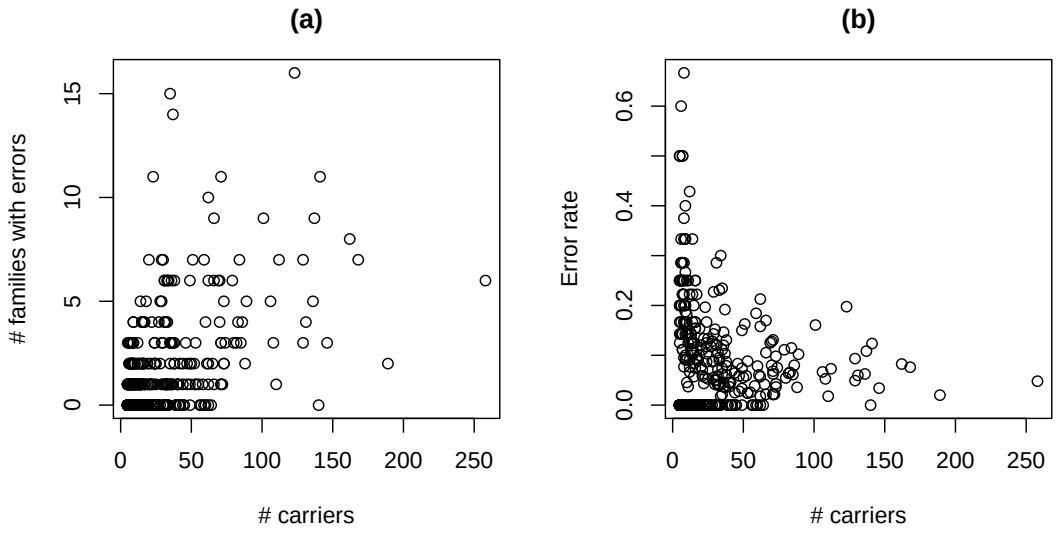


Figure 4.7: Mendelian errors in nuclear families for 446 CNPs. Each dot corresponds a CNP. (a) The number of nuclear families with Mendelian errors versus the number of CNV carriers; (b) The error rate versus the number of CNP carriers, where the error rate is defined as the ratio of the number of families with Mendelian errors among the number of families with CNP carriers.

members as unrelated. Both models allow users to specify prior information about allele frequency and invoke a Bayesian analysis. In our analysis, the prior frequency of the five alleles is set as $(0.04, 0.90, 0.04, 0.01, 0.01)$, which favors normal copy allele C1 and suppresses high copy allele C3 and C4 rarely seen. Note that the prior weights for deletion allele C0 and most common duplication allele C2 are set as equal. We used both models for frequency analysis and compared their estimates.

Since the estimated frequencies of C3 and C4 are at least one order of magnitude lower than C0 and C2, we focus on the latter two and take them as surrogates of deletion and duplication. Figure 4.8 (a) and (b) present a comparison of frequency estimation between two methods for each of deletion and duplication alleles. The pedigree structure does not substantially alter the estimation when allele frequency is relatively high (> 0.01). However, at the lower end (frequency < 0.01), using pedigree information

results in increased frequency estimation for alleles different from C1. Following is our current explanation for this phenomena: CN=2 is typically interpreted as resulting from genotype C1/C1 in unrelated individuals, as this is by far the genotype with the highest probability given copy number 2. However, within a pedigree with CNV carriers, it is quite possible that the genotype C0/C2 might have higher conditional probability. The allele frequency approach based on pedigree structure captures this. It has to be noted that some of the CN=2 phenotypes that need to be interpreted as genotype C0/C2 might in fact represent miscalls of the actual copy numbers of these subjects (or their relatives) so that this analysis can also be used to further refine CNV calls. Figure 4.8 (c) and (d) summarize the estimated frequency of C0 and C2 using pedigree information. The median frequency of deletion allele is three times duplication allele, reflecting the difficulty in detecting duplications based on our data.

4.3.3 Characteristics of detected CNPs

To explore the functional impact of CNPs on human genome, we looked at the abundance of 446 detected CNPs with respect to some genomic features, such as segmental duplications, genes and exons. The information of segmental duplications [Bailey et al. (2002)] and genes (including exons) [Pruitt et al. (2005)] were retrieved from UCSC Genome Browser [Fujita et al. (2011); Karolchik et al. (2004)]. We considered the overlap between detected CNP regions and feature regions. The significance of overlap is evaluated based on a simple test, where the nominal p-value is generated from the upper tail probability of $Binomial(n, p)$. n is the number of CNPs under consideration (446 here), which are assumed to be uniformly distributed on the genome. p is the proportion of the genome occupied a certain genomic feature. The results are summarized in Table 4.7. Consistently with previous findings, these CNPs are enriched in genomic regions with segmental duplications (p-value = 9.8×10^{-72}), and less enriched in gene regions (p-value = 8.2×10^{-2}) [Hastings et al. (2009)]. However, the fact that the interaction between CNP and exon region is extremely significant based on our data (p-value =

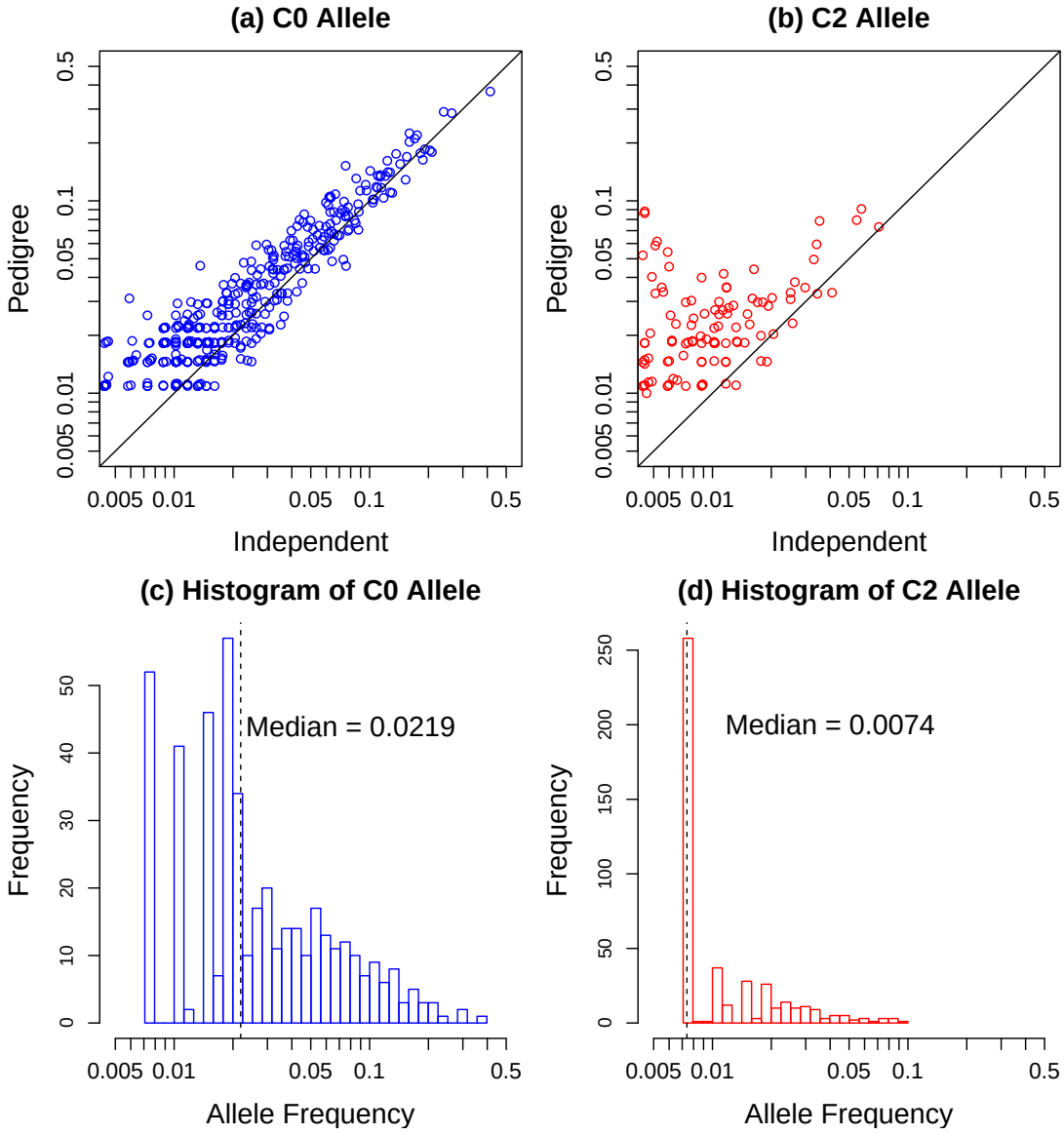


Figure 4.8: Frequency estimates of deletion allele (C0) and duplication allele (C2) for 446 CNPs. Frequencies are displayed on log scale. (a) and (b) contrast the frequency estimate based on pedigree information with the estimate treating subjects as independent for each of C0 and C2 alleles. The solid line indicates $y = x$. (c) and (d) are histograms of the frequency estimates using pedigree information. Dashed line marks the median frequency.

2.8×10^{-99}) is counter-intuitive. It might be possibly due to inaccurate specification of CNP boundaries. To further explore the overlap of CNP with exon, the CNP regions are broken down into three categories: those containing deletions only, duplication only and mixture of both (see Table 4.7). The CNP regions within which we observe duplications (Dup and Mix) have considerably larger percentage of overlap with exon than the regions containing deletions only (Del). It makes sense that extra copy of coding materials may supply additional redundancy for the functions of the genes [Hastings et al. (2009)]. We are further investigating these results.

It is known that SNPs have different frequencies across different populations. Here we investigated CNPs with this regard. Each of the 446 CNPs allows us to divide the sample into two subsets: the individuals carrying the CNP and those that do not. We contrast these two sets with respect to their belonging to CR versus CO populations and with respect to the reconstructed ancestries with the CNP region. More specifically, for the former purpose, one can construct a 2-by-2 contingency table for each CNP, with rows being CNP carriers versus non-carriers and columns being CR versus CO. For the latter purpose, one may construct a 2-by-3 contingency table for each CNP, with row being CNP carriers versus non-carriers and columns being the three ancestral components: Native American (N), European (E) and African (A). For example, a CNP carrier with the reconstructed local ancestry of N/E for a particular CNP region will have one count for each of the N and E cells in the table. Fisher's exact test was performed for each CNP. It is worth noting that the locus-specific ancestral labels were also subject to Mendelian error checking and only ancestral labels that showed consistent inheritance pattern were used for the analysis. At this stage, we relied only on the 67 "unrelated" subjects (see Section 4.1). No CNP shows significant difference in the frequencies of CNP carriers between CR and CO populations at a significance level with Bonferroni correction for multiple testing ($0.05/446 = 1.1 \times 10^{-4}$), while in differentiation of local ancestral components, only one CNP (on Chromosome 22q13.1) shows significance at the same level. However, one needs to observe that we have

Table 4.7: Summary statistics for enrichment of 446 detected CNPs with respect to different genomic features. The 22 autosomes have a total length of 2867.73 Mb. Listed are the number of each genomic feature, their total size, the percentage of the autosome occupied by each feature, the number and percentage of CNPs overlapping with each feature. In consideration of the overlap of CNP with exon, the CNP regions are further broken down into three categories: deletion, duplication and mixture of both. SD: segmental duplication; Del: deletion; Dup: duplication; Mix: mixture of deletion and duplication.

Genomic feature	Number	Total size (Mb)	% of the autosome	# overlapping CNPs	% overlapping CNPs
CNP	446	34.53	1.20%	–	–
SD	7264	130.45	4.55%	135	30.27%
Gene	18981	1157.87	40.38%	194	43.50%
Exon	207261	67.02	2.34%	128	28.70%
Del	309	–	–	46	14.89%
Mix	99	–	–	62	62.63%
Dup	38	–	–	20	52.63%

limited power when carrying out this test using only 67 subjects.

In addition to the 446 CNPs, we also considered the 239 detected CNPs which do not overlap with the reference CNPs compiled from DGV and dbVar (see Section 4.3.2). For each of the three ancestral populations, we calculated the average ancestral proportion over CNP carriers and non-carriers respectively, with respect to each CNP. The calculation is based on the data of 67 "unrelated" individuals (see Section 4.2.1). Figure 4.9 shows the difference in ancestral proportions between CNP carriers and non-carriers. The estimated ancestral proportions for CNP carriers generally have larger variance than non-carriers, largely due to limited sample size. The 239 non-overlapping CNPs seem to have a similar pattern in the distribution of ancestral proportions as the 446 CNPs. The observations do not change substantially when the comparison is extended to all 455 individuals (see Figure 4.10).

We may anticipate that the power of CNP in differentiating genetic backgrounds will increase as genotype data can be obtained from additional subjects. Moreover, if we combine the information from multiple sites of CNP for the admixture analysis, we expect to have better power.

4.4 Conclusions

We conducted a survey on the pedigree data of CR and CO populations in terms of admixture analysis and CNV analysis. These analyses illustrate the usage of proposed CNV detection method in a typical genetic study, especially in detection of copy number polymorphisms. The resulting list of CNPs together with locus specific ancestral inference provide useful information for future effort in identifying BP-related genetic variants and the interpretation of such findings.

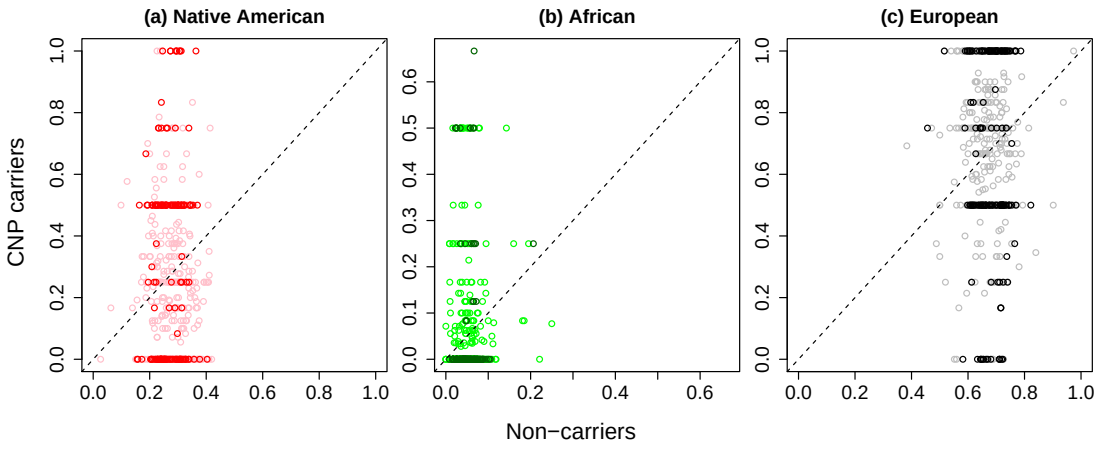


Figure 4.9: Difference in ancestral proportions between CNP carriers and non-carriers based on the subset of 67 "unrelated" subjects. For each of the three ancestral populations, the ancestral proportion for a CNP is averaged over CNP carriers and non-carriers respectively. A dot represents a CNP with lighter color indicating the ones that overlap with known CNPs and darker color indicating those having no overlap. The dashed line indicates $y = x$.

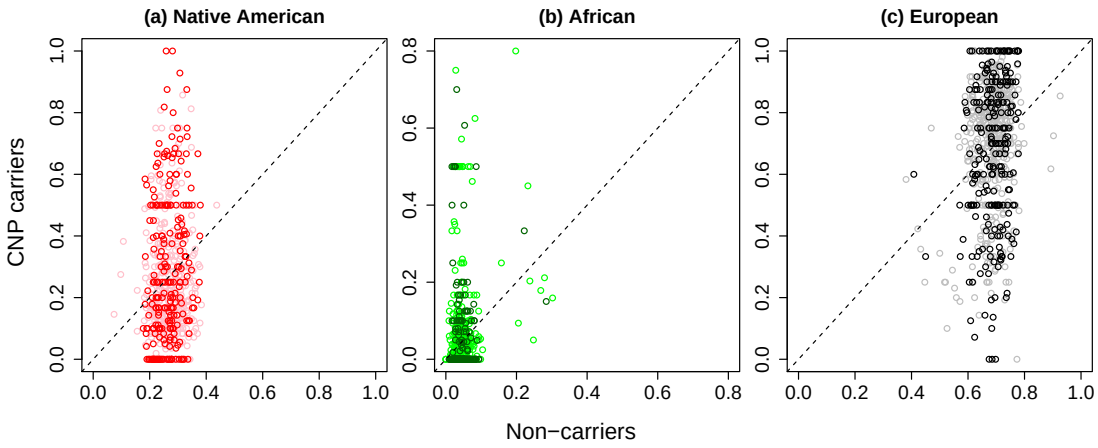


Figure 4.10: Difference in ancestral proportions between CNP carriers and non-carriers based on all 455 subjects. For each of the three ancestral populations, the ancestral proportion for a CNP is averaged over CNP carriers and non-carriers respectively. A dot represents a CNP with lighter color indicating the ones that overlap with known CNPs and darker color indicating those having no overlap. The dashed line indicates $y = x$.

BIBLIOGRAPHY

D. Albertson and D. Pinkel. Genomic microarrays in human genetic disease and cancer. *Human Molecular Genetics*, 12(Review Issue 2):R145–R152, 2003.

D. G. Albertson, C. Collins, F. McCormick, and J. W. Gray. Chromosome aberrations in solid tumors. *Nature Genetics*, 34(4):369–376, 2003.

D. H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664, 2009.

J. A. Bailey, Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, S. Schwartz, M. D. Adams, E. W. Myers, P. W. Li, and E. E. Eichler. Recent segmental duplications in the human genome. *Science*, 297(5583):1003–1007, 2002.

M. T. Barrett, A. Scheffer, A. Ben-Dor, N. Sampas, D. Lipson, R. Kincaid, P. Tsang, B. Curry, K. Baird, P. Meltzer, Z. Yakhini, L. Bruhn, and S. Laderman. Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 101(51):17765, 2004.

H. Bengtsson, R. Irizarry, B. Carvalho, and T. Speed. Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics*, 24(6):759–767, 2008.

H. Bengtsson, P. Wirapati, and T. P. Speed. A single-array preprocessing method for estimating full-resolution raw copy numbers from all affymetrix genotyping arrays including genomewidesnp 5 & 6. *Bioinformatics*, 25(17):2149–2156, 2009.

H. Bengtsson, P. Neuvial, and T. Speed. Tumorboost: Normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays. *BMC Bioinformatics*, 11(1):245, 2010.

R. W. Bentley, J. Pearson, R. B. Gearry, M. L. Barclay, C. McKinney, T. R. Merriman, and R. L. Roberts. Association of higher defb4 genomic copy number with crohn's disease. *The American Journal of Gastroenterology*, 105(2):354–359, 2009.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

G. R. Bignell, J. Huang, J. Greshock, S. Watt, A. Butler, S. West, M. Grigorova, K. W. Jones, W. Wei, M. R. Stratton, P. Andrew Futreal, B. Weber, M. H. Shapero, and R. Wooster. High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Research*, 14(2):287–295, 2004.

J. M. Bioucas-Diaa, M. A. T. Figueiredo, and J. P. Oliveira. Adaptive total variation image deconvolution: A majorization-minimization approach. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*. Toulouse, France, 2006.

K. Bleakley and J. P. Vert. The group fused lasso for multiple change-point detection. *Arxiv preprint arXiv:1106.4199*, 2011.

F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

C. D. Campbell, N. Sampas, A. Tsalenko, P. H. Sudmant, J. M. Kidd, M. Malig, T. H. Vu, L. Vives, P. Tsang, L. Bruhn, and E. E. Eichler. Population-genetic properties of differentiated human copy-number polymorphisms. *The American Journal of Human Genetics*, 88(3):317–332, 2011.

E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

E. J. Candès and Y. Plan. Near-ideal model selection by ℓ_1 minimization. *The Annals of Statistics*, 37(5A):2145–2177, 2009.

L. G. Carvajal-Carmona, R. Ophoff, S. Service, J. Hartiala, J. Molina, P. Leon, J. Ospina, G. Bedoya, N. Freimer, and A. Ruiz-Linares. Genetic demography of antioquia (colombia) and the central valley of costa rica. *Human Genetics*, 112(5):534–541, 2003.

B. Carvalho, H. Bengtsson, T. P. Speed, and R. A. Irizarry. Exploration, normalization, and genotype calls of high-density oligonucleotide snp array data. *Biostatistics*, 8 (2):485–499, 2007.

T. F. Chan and J. Shen. Mathematical models for local nontexture inpainting. *SIAM Journal on Applied Mathematics*, 62:1019–1043, 2002.

H. Chen, H. Xing, and N. R. Zhang. Estimation of parent specific DNA copy number in tumors using high-density genotyping arrays. *PLoS Computational Biology*, 7(1): e1001060, 2011a.

Y. Chen, Y. J. Liu, Y. F. Pei, T. L. Yang, F. Y. Deng, X. G. Liu, D. Y. Li, and H. W. Deng. Copy number variations at the prader–willi syndrome region on chromosome 15 and associations with obesity in whites. *Obesity*, 19(6):1229–1234, 2011b.

S. Colella, C. Yau, J. M. Taylor, G. Mirza, H. Butler, P. Clouston, A. S. Bassett, A. Seller, C. C. Holmes, and J. Ragoussis. QuantiSNP: An objective Bayes hidden-Markov model to detect and accurately map copy number variation using snp genotyping data. *Nucleic Acids Research*, 35(6):2013–2025, 2007.

D. F. Conrad, D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C. H. Ihm, K. Kristiansson, D. G. MacArthur, J. R. MacDonald, I. Onyiah, A. W. C. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, T. W. T. C. C. Consortium, C. Tyler-Smith, N. P. Carter, C. Lee, S. W. Scherer, and M. E. Hurles. Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–712, 2009.

S. D. Conte and C. deBoor. *Elementary Numerical Analysis*. McGraw-Hill, New York, 1972.

G. M. Cooper, T. Zerr, J. M. Kidd, E. E. Eichler, and D. A. Nickerson. Systematic assessment of copy number variant detection via genome-wide snp genotyping. *Nature Genetics*, 40(10):1199–1203, 2008.

A. E. Dellinger, S. M. Saw, L. K. Goh, M. Seielstad, T. L. Young, and Y. J. Li. Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nucleic Acids Research*, 38(9):e105, 2010.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1):1–38, 1977.

S. J. Diskin, M. Li, C. Hou, S. Yang, J. Glessner, H. Hakonarson, M. Bucan, J. M. Maris, and K. Wang. Adjustment of genomic waves in signal intensities from whole-genome snp genotyping platforms. *Nucleic Acids Research*, 36(19):e126, 2008.

D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

B. Efron and N. R. Zhang. False discovery rates and copy number variation. *Biometrika*, 98(2):251–271, 2011.

D. Falush, M. Stephens, and J. K. Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587, 2003.

M. Fanciulli, P. J. Norsworthy, E. Petretto, R. Dong, L. Harper, L. Kamesh, J. M. Heward, S. C. L. Gough, A. de Smith, A. I. F. Blakemore, P. Froguel, C. J. Owen, S. H. S. Pearce, L. Teixeira, L. Guillevin, D. S. Cunninghame Graham, C. D. Pusey,

H. Terence Cook, T. J. Vyse, and T. J. Aitman. Fcgr3b copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nature genetics*, 39(6):721–723, 2007.

L. Feuk, A. R. Carson, and S. W. Scherer. Structural variation in the human genome. *Nature Reviews Genetics*, 7(2):85–97, 2006.

J. L. Freeman, G. H. Perry, L. Feuk, R. Redon, S. A. McCarroll, D. M. Altshuler, H. Aburatani, K. W. Jones, C. Tyler-Smith, M. E. Hurles, N. P. Carter, S. W. Scherer, and C. Lee. Copy number variation: New insights in genome diversity. *Genome research*, 16(8):949–961, 2006.

J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. *Arxiv preprint arXiv:1001.0736*, 2010.

P. A. Fujita, B. Rhead, A. S. Zweig, A. S. Hinrichs, D. Karolchik, M. S. Cline, M. Goldman, G. P. Barber, H. Clawson, A. Coelho, M. Diekhans, T. R. Dreszer, B. M. Giardine, R. A. Harte, J. Hillman-Jackson, F. Hsu, V. Kirkup, R. M. Kuhn, K. Learned, C. H. Li, L. R. Meyer, A. Pohl, B. J. Raney, K. R. Rosenbloom, K. E. Smith, D. Haussler, and W. J. Kent. The ucsc genome browser database: update 2011. *Nucleic Acids Research*, 39(Suppl 1):D876–D882, 2011.

D. B. Goldstein and G. L. Cavalleri. Understanding human diversity. *Nature*, 437 (7063):1241–1242, 2005.

P. J. Hastings, J. R. Lupski, S. M. Rosenberg, and G. Ira. Mechanisms of change in gene copy number. *Nature Reviews Genetics*, 10(8):551–564, 2009.

H. Hoefling. A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19(4):984–1006, 2010.

E. J. Hollox, U. Huffmeier, P. L. J. M. Zeeuwen, R. Palla, J. Lascorz, D. Rodijk-Olthuis, P. C. M. van de Kerkhof, H. Traupe, G. de Jongh, M. den Heijer, A. Reis, J. A. L. Armour, and J. Schalkwijk. Psoriasis is associated with increased β-defensin genomic copy number. *Nature Genetics*, 40(1):23–25, 2008.

J. Huang, W. Wei, J. Zhang, G. Liu, G. R. Bignell, M. R. Stratton, P. A. Futreal, R. Wooster, K. W. Jones, and M. H. Shapero. Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Human Genomics*, 1(4): 287–299, 2004.

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

A. Ingason, D. Rujescu, S. Cichon, E. Sigurdsson, T. Sigmundsson, O. P. H. Pietiläinen, J. E. Buizer-Voskamp, E. Strengman, C. Francks, P. Muglia, A. Gylfason, O. Gustafsson, P. I. Olason, S. Steinberg, T. Hansen, K. D. Jakobsen, H. B. Rasmussen, I. Giegling, H. J. Möller, A. Hartmann, C. Crombie, G. Fraser, N. Walker, J. Lonnqvist, J. Suvisaari, A. Tuulio-Henriksson, E. Bramon, L. A. Kiemeney, B. Franke, R. Murray, E. Vassos, T. Toulopoulou, T. W. Mühleisen, S. Tosato, M. Ruggeri, S. Djurovic, O. A. Andreassen, Z. Zhang, T. Werge, R. A. Ophoff, GROUP Investigators, M. Rietschel, M. M. Nöthen, H. Petursson, H. Stefansson, L. Peltonen, D. Collier, K. Stefansson, and D. St Clair. Copy number variations of chromosome 16p13.1 region associated with schizophrenia. *Molecular Psychiatry*, 16(1):17–25, 2009.

M. Jakobsson, S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere, H. C. Fung, Z. A. Szpiech, J. H. Degnan, K. Wang, R. Guerreiro, J. M. Bras, J. C. Schymick, D. G. Hernandez, B. J. Traynor, J. Simon-Sanchez, M. Matarin, A. Britton, J. van de Leemput, I. Rafferty, M. Bucan, H. M. Cann, J. A. Hardy, N. A. Rosenberg, and A. B.

Singleton. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, 451(7181):998–1003, 2008.

I. Jarick, C. I. G. Vogel, S. Scherag, H. Schäfer, J. Hebebrand, A. Hinney, and A. Scherag. Novel common copy number variation for early onset extreme obesity on chromosome 11q11 identified by a genome-wide analysis. *Human Molecular Genetics*, 20(4):840–852, 2011.

D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, and W. J. Kent. The ucsc table browser data retrieval tool. *Nucleic Acids Research*, 32(Suppl 1):D493–D496, 2004.

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

J. O. Korbel, A. E. Urban, J. P. Affourtit, B. Godwin, F. Grubert, J. F. Simons, P. M. Kim, D. Palejev, N. J. Carriero, L. Du, B. E. Taillon, Z. Chen, A. Tanzer, A. C. Eugenia Saunders, J. Chi, F. Yang, N. P. Carter, M. E. Hurles, S. M. Weissman, T. T. Harkins, M. B. Gerstein, M. Egholm, and M. Snyder. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318(5849):420–426, 2007.

J. M. Korn, F. G. Kuruvilla, S. A. McCarroll, A. Wysoker, J. Nemesh, S. Cawley, E. Hubbell, J. Veitch, P. J. Collins, K. Darvishi, C. Lee, M. M. Nizzari, S. B. Gabriel,

S. Purcell, M. J. Daly, and D. Altshuler. Integrated genotype calling and association analysis of snps, common copy number polymorphisms and rare cnvs. *Nature Genetics*, 40(10):1253–1260, 2008.

T. Lai, H. Xing, and N. Zhang. Stochastic segmentation models for array-based comparative genomic hybridization data analysis. *Biostatistics*, 9(2):290–307, 2008.

W. R. Lai, M. D. Johnson, R. Kucherlapati, and P. J. Park. Comparative analysis of algorithms for identifying amplifications and deletions in array cgh data. *Bioinformatics*, 21(19):3763–3770, 2005.

K. Lange. *Optimization*. Springer, New York, 2004.

K. Lange, R. Cantor, S. Horvath, M. Perola, C. Sabatti, J. Sinsheimer, and E. Sobel. Mendel version 4.0: A complete package for the exact genetic analysis of discrete traits in pedigree and population data sets. *The American Journal of Human Genetics*, 69(Suppl 1):504, 2001.

K. Lange, J. S. Sinsheimer, and E. Sobel. Association testing with mendel. *Genetic Epidemiology*, 29(1):36–50, 2005.

Y. Li and J. Zhu. Analysis of array cgh data for cancer studies using fused quantile regression. *Bioinformatics*, 23(18):2470–2476, 2007.

R. Lucito, J. Healy, J. Alexander, A. Reiner, D. Esposito, M. Chi, L. Rodgers, A. Brady, J. Sebat, J. Troge, J. A. West, S. Rostan, K. C. Q. Nguyen, S. Powers, K. Q. Ye, A. Olshen, E. Venkatraman, L. Norton, and M. Wigler. Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome research*, 13(10):2291–2305, 2003.

D. Malhotra, S. McCarthy, J. J. Michaelson, V. Vacic, K. E. Burdick, S. Yoon, S. Cichon, A. Corvin, S. Gary, E. S. Gershon, M. Gill, M. Karayiorgou, J. R. Kelsoe, O. Krastoshevsky, V. Krause, E. Leibenluft, D. L. Levy, V. Makarov, A. Bhandari,

A. K. Malhotra, F. J. McMahon, M. M. Nöthen, J. B. Potash, M. Rietschel, T. G. Schulze, and J. Sebat. High frequencies of de novo cnvs in bipolar disorder and schizophrenia. *Neuron*, 72(6):951–963, 2011.

J. C. Marioni, N. P. Thorne, A. Valsesia, T. Fitzgerald, R. Redon, H. Fiegler, T. D. Andrews, B. E. Stranger, A. G. Lynch, E. T. Dermitzakis, N. P. Carter, S. Tavaré, and M. E. Hurles. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biology*, 8(10): R228, 2007.

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

R. E. Mills, K. Walter, C. Stewart, R. E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S. C. Yoon, K. Ye, R. K. Cheetham, A. Chinwalla, D. F. Conrad, Y. Fu, F. Grubert, I. Hajirasouliha, F. Hormozdiari, L. M. Iakoucheva, Z. Iqbal, S. Kang, J. M. Kidd, M. K. Konkel, J. Korn, E. Khurana, D. Kural, H. Y. K. Lam, J. Leng, R. Li, Y. Li, C.-Y. Lin, R. Luo, X. J. Mu, J. Nemesh, H. E. Peckham, T. Rausch, A. Scally, X. Shi, M. P. Stromberg, A. M. Stutz, A. E. Urban, J. A. Walker, J. Wu, Y. Zhang, Z. D. Zhang, M. A. Batzer, L. Ding, G. T. Marth, G. McVean, J. Sebat, M. Snyder, J. Wang, K. Ye, E. E. Eichler, M. B. Gerstein, M. E. Hurles, C. Lee, S. A. McCarroll, J. O. Korbel, and . G. Project. Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332):59–65, 2011.

M. Molokhia, M. Fanciulli, E. Petretto, A. L. Patrick, P. McKeigue, A. L. Roberts, T. J. Vyse, and T. J. Aitman. Fcgr3b copy number variation is associated with systemic lupus erythematosus risk in afro-caribbeans. *Rheumatology*, 50(7):1206–1210, 2011.

S. Negahban, P. Ravikmuar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *The Neural Information Processing Systems Conference (NIPS'09)*. Vancouver, Canada, 2009.

P. Neuvial, H. Bengtsson, and T. P. Speed. Statistical analysis of single nucleotide polymorphism microarrays in cancer studies. *Handbook of Statistical Bioinformatics*, pages 225–255, 2011.

M. A. Newton and Y. Lee. Inferring the location and effect of tumor suppressor genes by instability-selection modeling of allelic-loss data. *Biometrics*, 56(4):1088–1097, 2000.

G. Nowak, T. Hastie, J. R. Pollack, and R. Tibshirani. A fused lasso latent feature model for analyzing multi-sample acgh data. *Biostatistics*, 12(4):776–791, 2011.

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

M. Ortiz-Estevez, H. Bengtsson, and A. Rubio. ACNE: a summarization method to estimate allele-specific copy numbers for affymetrix snp arrays. *Bioinformatics*, 26 (15):1827–1833, 2010.

M. Ortiz-Estevez, A. Aramburu, H. Bengtsson, P. Neuvial, and A. Rubio. CalMaTe: A method and software to improve allele-specific copy number of snp arrays for downstream segmentation. *Bioinformatics*, Online Access, 2012.

I. Ostrovnaya, A. B. Olshen, V. E. Seshan, I. Orlow, D. G. Albertson, and C. B. Begg. A metastasis or a second independent cancer? Evaluating the clonal origin of tumors using array copy number data. *Statistics in medicine*, 29(15):1608–1621, 2010.

B. Paşaniuc, S. Sankararaman, G. Kimmel, and E. Halperin. Inference of locus-specific ancestry in closely related populations. *Bioinformatics*, 25(12):i213–i221, 2009.

N. Patterson, N. Hattangadi, B. Lane, K. E. Lohmueller, D. A. Hafler, J. R. Oksenberg, S. L. Hauser, M. W. Smith, S. J. O'Brien, D. Altshuler, M. J. Daly, and D. Reich. Methods for high-density admixture mapping of disease genes. *The American Journal of Human Genetics*, 74(5):979–1000, 2004.

D. A. Peiffer, J. M. Le, F. J. Steemers, W. Chang, T. Jenniges, F. Garcia, K. Haden, J. Li, C. A. Shaw, J. Belmont, S. W. Cheung, R. M. Shen, D. L. Barker, and K. L. Gunderson. High-resolution genomic profiling of chromosomal aberrations using infinium whole-genome genotyping. *Genome Research*, 16(9):1136–1148, 2006.

F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J. J. Daudin. A statistical approach for array cgh data analysis. *BMC Bioinformatics*, 6(1):27, 2005.

D. Pinkel and D. G. Albertson. Array comparative genomic hybridization and its applications in cancer. *Nature Genetics*, 37:S11–S17, 2005.

D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B. M. Ljung, J. W. Gray, and D. G. Albertson. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, 20(2):207–211, 1998.

D. Pinto, A. T. Pagnamenta, L. Klei, R. Anney, D. Merico, R. Regan, J. Conroy, T. R. Magalhaes, C. Correia, B. S. Abrahams, J. Almeida, E. Bacchelli, G. D. Bader, A. J. Bailey, G. Baird, A. Battaglia, T. Berney, N. Bolshakova, S. Bölte, P. F. Bolton, T. Bourgeron, S. Brennan, J. Brian, S. E. Bryson, A. R. Carson, G. Casallo, J. Casey, B. H. Y. Chung, L. Cochrane, C. Corsello, E. L. Crawford, A. Crossett, C. Cytrynbaum, G. Dawson, M. de Jonge, R. Delorme, I. Drmic, E. Duketis, F. Duque, A. Estes, P. Farrar, B. A. Fernandez, S. E. Folstein, E. Fombonne, C. M. Freitag, J. Gilbert, C. Gillberg, J. T. Glessner, J. Goldberg, A. Green, J. Green, S. J. Guter,

H. Hakonarson, E. A. Heron, M. Hill, R. Holt, J. L. Howe, G. Hughes, V. Hus, R. Igliozzi, C. Kim, S. M. Klauck, A. Kolevzon, O. Korvatska, V. Kustanovich, C. M. Lajonchere, J. A. Lamb, M. Laskawiec, M. Leboyer, A. Le Couteur, B. L. Leventhal, A. C. Lionel, X.-Q. Liu, C. Lord, L. Lotspeich, S. C. Lund, E. Maestrini, W. Mahoney, C. Mantoulan, C. R. Marshall, H. McConachie, C. J. McDougle, J. McGrath, W. M. McMahon, A. Merikangas, O. Migita, N. J. Minshew, G. K. Mirza, J. Munson, S. F. Nelson, C. Noakes, A. Noor, G. Nygren, G. Oliveira, K. Papanikolaou, J. R. Parr, B. Parrini, T. Paton, A. Pickles, M. Pilorge, J. Piven, C. P. Ponting, D. J. Posey, A. Poustka, F. Poustka, A. Prasad, J. Ragoussis, K. Renshaw, J. Rickaby, W. Roberts, K. Roeder, B. Roge, M. L. Rutter, L. J. Bierut, J. P. Rice, J. Salt, K. Sansom, D. Sato, R. Segurado, A. F. Sequeira, L. Senman, N. Shah, V. C. Sheffield, L. Soorya, I. Sousa, O. Stein, N. Sykes, V. Stoppioni, C. Strawbridge, R. Tancredi, K. Tansey, B. Thiruvahindrapduram, A. P. Thompson, S. Thomson, A. Tryfon, J. Tsiantis, H. Van Engeland, J. B. Vincent, F. Volkmar, S. Wallace, K. Wang, Z. Wang, T. H. Wassink, C. Webber, R. Weksberg, K. Wing, K. Wittemeyer, S. Wood, J. Wu, B. L. Yaspan, D. Zurawiecki, L. Zwaigenbaum, J. D. Buxbaum, R. M. Cantor, E. H. Cook, H. Coon, M. L. Cuccaro, B. Devlin, S. Ennis, L. Gallagher, D. H. Geschwind, M. Gill, J. L. Haines, J. Hallmayer, J. Miller, A. P. Monaco, J. I. Nurnberger Jr, A. D. Paterson, M. A. Pericak-Vance, G. D. Schellenberg, P. Szatmari, A. M. Vicente, V. J. Vieland, E. M. Wijsman, S. W. Scherer, S. J. S., and C. Betancur. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, 466(7304):368–372, 2010.

D. Pinto, K. Darvishi, X. Shi, D. Rajan, D. Rigler, T. Fitzgerald, A. C. Lionel, B. Thiruvahindrapuram, J. R. MacDonald, R. Mills, A. Prasad, K. Noonan, S. Gribble, E. Prigmore, P. K. Donahoe, R. S. Smith, J. H. Park, M. E. Hurles, N. P. Carter, C. Lee, S. W. Scherer, and L. Feuk. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature Biotechnology*, 29(6):512–520, 2011.

J. R. Pollack, C. M. Perou, A. A. Alizadeh, M. B. Eisen, A. Pergamenschikov, C. F. Williams, S. S. Jeffrey, D. Botstein, and P. O. Brown. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics*, 23(1):41–46, 1999.

J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

K. D. Pruitt, T. Tatusova, and D. R. Maglott. Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 33(Suppl 1):D501–D504, 2005.

L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. Gonzalez, M. Gratacos, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer, and M. E. Hurles. Global variation in copy number in the human genome. *Nature*, 444(7118):444–454, 2006.

A. Rinaldo. Properties and refinements of the fused lasso. *The Annals of Statistics*, 37 (5B):2922–2952, 2009.

N. Risch. Mapping genes for complex disease using association studies with recently admixed populations. *The American Journal of Human Genetics*, Suppl 51:13, 1992.

L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60(1-4):259–268, 1992.

S. Sankararaman, G. Kimmel, E. Halperin, and M. I. Jordan. On the inference of ancestries in admixed populations. *Genome Research*, 18(4):668–675, 2008a.

S. Sankararaman, S. Sridhar, G. Kimmel, and E. Halperin. Estimating local ancestry in admixed populations. *The American Journal of Human Genetics*, 82(2):290–303, 2008b.

R. B. Scharpf, G. Parmigiani, J. Pevsner, and I. Ruczinski. Hidden Markov models for the assessment of chromosomal alterations using high throughput snp arrays. *The Annals of Applied Statistics*, 2(2):687–713, 2008.

R. B. Scharpf, M. E. Irizarry R. A., Ritchie, B. Carvalho, and I. Ruczinski. Using the r package crlmm for genotyping and copy number estimation. *Journal of Statistical Software*, 40(i12):1–32, 2011a.

R. B. Scharpf, I. Ruczinski, B. Carvalho, B. Doan, A. Chakravarti, and R. A. Irizarry. A multilevel model to address batch effects in copy number estimation using snp arrays. *Biostatistics*, 12(1):33–50, 2011b.

S. Scherer, C. Lee, E. Birney, D. Altshuler, E. Eichler, N. Carter, M. Hurles, and L. Feuk. Challenges and standards in integrating surveys of structural variation. *Nature Genetics*, 39:S7–S15, 2007.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978.

J. Sebat. Major changes in our DNA lead to major changes in our thinking. *Nature Genetics*, 39:S3–S5, 2007.

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

S. Service, J. DeYoung, M. Karayiorgou, J. L. Roos, H. Pretorious, G. Bedoya, J. Ospina, A. Ruiz-Linares, A. Macedo, J. A. Palha, P. Heutink, Y. Aulchenko, B. Oostra, C. van Duijn, M.-R. Jarvelin, T. Varilo, L. Peddle, P. Rahman, G. Piras, M. Monne, S. Murray, L. Galver, L. Peltonen, C. Sabatti, A. Collins, and N. Freimer. Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nature Genetics*, 38(5):556–560, 2006.

D. O. Siegmund, B. Yakir, and N. R. Zhang. Detecting simultaneous intervals in aligned sequences. *The Annals of Applied Statistics*, 5(2A):645–668, 2011.

A. M. Snijders, N. Nowak, R. Segraves, S. Blackwood, N. Brown, J. Conroy, G. Hamilton, A. K. Hindle, B. Huey, K. Kimura, S. Law, K. Myambo, J. Palmer, B. Ylstra, J. P. Yue, J. W. Gray, A. N. Jain, D. Pinkel, and D. G. Albertson. Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics*, 29 (3):263–264, 2001.

E. Sobel, J. C. Papp, and K. Lange. Detection and integration of genotyping errors in statistical genetics. *The American Journal of Human Genetics*, 70(2):496–508, 2002.

J. Staaf, D. Lindgren, J. Vallon-Christersson, A. Isaksson, H. Göransson, G. Juliusson, R. Rosenquist, M. Höglund, A. Borg, and M. Ringnér. Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome snp arrays. *Genome Biology*, 9(9):R136, 2008.

H. Stefansson, D. Rujescu, S. Cichon, O. P. H. Pietiläinen, A. Ingason, S. Steinberg, R. Fossdal, E. Sigurdsson, T. Sigmundsson, J. E. Buizer-Voskamp, T. Hansen, K. D. Jakobsen, P. Muglia, C. Francks, P. M. Matthews, A. Gylfason, B. V. Halldorsson, D. Gudbjartsson, T. E. Thorgeirsson, A. Sigurdsson, A. Jonasdottir, A. Jonasdottir, A. Bjornsson, S. Mattiasdottir, T. Blondal, M. Haraldsson, B. B. Magnusdottir, I. Giegling, H.-J. Möller, A. Hartmann, K. V. Shianna, D. Ge, A. C. Need, C. Crombie, G. Fraser, N. Walker, J. Lonnqvist, J. Suvisaari, A. Tuulio-Henriksson, T. Paunio, T. Toulopoulou, E. Bramon, M. Di Forti, R. Murray, M. Ruggeri, E. Vassos,

S. Tosato, M. Walshe, T. Li, C. Vasilescu, T. W. Mühleisen, A. G. Wang, H. Ullum, S. Djurovic, I. Melle, J. Olesen, L. A. Kiemeney, B. Franke, Genetic Risk and Outcome in Psychosis (GROUP), C. Sabatti, N. B. Freimer, J. R. Gulcher, U. Thorsteinsdottir, A. Kong, O. A. Andreassen, R. A. Ophoff, A. Georgi, M. Rietschel, T. Werge, H. Petursson, D. B. Goldstein, M. M. Nöthen, L. Peltonen, D. A. Collier, D. St Clair, and K. Stefansson. Large recurrent microdeletions associated with schizophrenia. *Nature*, 455(7210):232–236, 2008.

W. Sun, F. A. Wright, Z. Tang, S. H. Nordgard, P. Van Loo, T. Yu, V. N. Kristensen, and C. M. Perou. Integrated study of copy number states and genotype calls using high-density snp arrays. *Nucleic Acids Research*, 37(16):5365–5377, 2009.

H. Tang, J. Peng, P. Wang, and N. J. Risch. Estimation of individual admixture: analytical and study design considerations. *Genetic Epidemiology*, 28(4):289–301, 2005.

H. Tang, M. Coram, P. Wang, X. Zhu, and N. Risch. Reconstructing genetic ancestry blocks in admixed individuals. *The American Journal of Human Genetics*, 79(1): 1–12, 2006.

The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, 2010.

The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 2005.

The International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*, 455(7210):237–241, 2008.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 1996.

R. Tibshirani and P. Wang. Spatial smoothing and hot spot detection for cgh data using the fused lasso. *Biostatistics*, 9(1):18–29, 2008.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 67(1):91–108, 2005.

R. J. Tibshirani and J. Taylor. The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371, 2011.

E. Tuzun, A. J. Sharp, J. A. Bailey, R. Kaul, V. A. Morrison, L. M. Pertz, E. Haugen, H. Hayden, D. Albertson, D. Pinkel, M. V. Olson, and E. E. Eichler. Fine-scale structural variation of the human genome. *Nature Genetics*, 37(7):727–732, 2005.

E. S. Venkatraman and A. B. Olshen. A faster circular binary segmentation algorithm for the analysis of array cgh data. *Bioinformatics*, 23(6):657–663, 2007.

T. Vrijenhoek, J. E. Buizer-Voskamp, I. van der Stelt, E. Strengman, Genetic Risk and Outcome in Psychosis (GROUP) Consortium, C. Sabatti, A. G. van Kessel, H. G. Brunner, R. A. Ophoff, and J. A. Veltman. Recurrent cnvs disrupt three candidate genes in schizophrenia patients. *The American Journal of Human Genetics*, 83(4): 504–510, 2008.

T. Walsh, J. M. McClellan, S. E. McCarthy, A. M. Addington, S. B. Pierce, G. M. Cooper, A. S. Nord, M. Kusenda, D. Malhotra, A. Bhandari, S. M. Stray, C. F. Rippey, P. Roccanova, V. Makarov, B. Lakshmi, R. L. Findling, L. Sikich, T. Stromberg, B. Merriman, N. Gogtay, P. Butler, K. Eckstrand, L. Noory, P. Gochman, R. Long, Z. Chen, S. Davis, C. Baker, E. E. Eichler, P. S. Meltzer, S. F. Nelson, A. B. Singleton, M. K. Lee, J. L. Rapoport, M.-C. King, and J. Sebat. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science's STKE*, 320(5875):539, 2008.

H. Wang, J. H. Veldink, H. Blauw, L. H. van den Berg, R. A. Ophoff, and C. Sabatti. Markov models for inferring copy number variations from genotype data on illumina platforms. *Human Heredity*, 68:1–22, 2009.

K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S. F. A. Grant, H. Hakonarson, and M. Bucan. 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, 2007.

K. Wang, Z. Chen, M. G. Tadesse, J. Glessner, S. F. A. Grant, H. Hakonarson, M. Bucan, and M. Li. Modeling genetic inheritance of copy number variations. *Nucleic Acids Research*, 36(21):e138, 2008.

H. Willenbrock and J. Fridlyand. A comparison study: applying segmentation to array cgh data for downstream analyses. *Bioinformatics*, 21(22):4084–4091, 2005.

T. T. Wu and K. Lange. Coordinate descent algorithm for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244, 2008.

T. L. Yang, Y. Guo, S. M. Li, S. K. Li, Q. Tian, Y. J. Liu, and H. W. Deng. Ethnic differentiation of copy number variation on chromosome 16p12.3 for association with obesity phenotypes in european and chinese populations. *International Journal of Obesity*, Online Access, 2012.

C. Yau, O. Papaspiliopoulos, G. O. Roberts, and C. Holmes. Bayesian non-parametric hidden Markov models with applications in genomics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):37–57, 2011.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

J. Zhang, L. Feuk, G. E. Duggan, R. Khaja, and S. W. Scherer. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenetic and Genome Research*, 115(3-4):205–214, 2006.

N. R. Zhang and D. O. Siegmund. A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63 (1):22–32, 2007.

N. R. Zhang, Y. Senbabaoglu, and J. Z. Li. Joint estimation of DNA copy number from multiple platforms. *Bioinfomatics*, 26(2):153–160, 2010a.

N. R. Zhang, D. O. Siegmund, H. Ji, and J. Z. Li. Detecting simultaneous changepoints in multiple sequences. *Biometrika*, 97(3):631–645, 2010b.

Z. Zhang, K. Lange, R. Ophoff, and C. Sabatti. Reconstructing DNA copy number by penalized estimation and imputation. *The Annals of Applied Statistics*, 4(4):1749–1773, 2010c.

H. Zhou and K. Lange. A path algorithm for constrained estimation. *Arxiv preprint arXiv:1103.3738*, 2011.

H. Zhou, M. E. Sehl, J. S. Sinsheimer, and K. Lange. Association screening of common and rare genetic variants by penalized regression. *Bioinformatics*, 26(19):2375–2382, 2010.