ABSTRACT

HARIANI, GUNJAN D. Application of Next Generation Sequencing Technologies to Pharmacogenomics. (Under the direction of Dr. Alison A. Motsinger-Reif).

Next Generation Sequencing (NGS) technologies have made it possible to assess genetic diversity at an unprecedented pace due to new levels of parallelization which allows for cheap and fast data turnaround. These technologies are making it possible for the researchers to discover previously uncharacterized variants and investigate their role with trait variability; which is an important question in the field of genetics. One hypothesis that has particularly gained interest is the role of low frequency variants (variants that are seen in less than 5% of the populations) in common diseases and traits (also known as the common disease rare variant or CDRV hypothesis).

One of the first things I did in this dissertation is to understand the rationale of pursuing the CDRV hypothesis. I then performed power comparisons for a few statistical methods that have been especially designed for testing this hypothesis and applied some of these tests to examine the role of low frequency variants in a candidate gene study (paternally expressed 3 – PEG3) for a pregnancy related disorder known as small for gestational age in 55 samples. No excess accumulation of low frequency exonic variants was found in infants with SGA when compared to controls; which leaves the role of genetic variants in PEG3 and SGA unclear.

The other part of this dissertation has focused on application of sequencing technologies as a tool for discovery in pharmacogenomics. We were interested in investigating the role of rare and common variants in chemotherapy drug response. We obtained sequencing data for 95 CEPH cell lines (cell lines of Utah residents with ancestry from northern and western Europe) with drug response phenotypes for 30 FDA approved chemotherapeutics. I invested time to set up the pipeline to process this sequencing data. After rigorous data quality checks, I generated variant calls for these cell lines which were then used for association testing with drug response. The findings from this study demonstrate the successful application of sequencing of cell line models for discovery in pharmacogenomics. The amount of time taken to set up the pipeline only reciprocate the

challenges of working with huge amounts of data and I conclude that NGS technology is not applicable for every study as the costs of analysis may out-benefit the cost of data generation.

Application of Next Generation Sequencing Technologies to Pharmacogenomics

by
Gunjan Hariani

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Bioinformatics

Raleigh, North Carolina

2012

APPROVED BY:

_____	_____
Alison A. Motsinger-Reif Committee Chair	Dahlia Nielsen
_____	_____
Eric Stone	Jorge Piedrahita

DEDICATION

I would like to dedicate this to my parents, Dhanraj and Sunita Hariani for their love, support and patience. Also, to my beautiful sister, Taruna Madupalli who cares for me deeply.

BIOGRAPHY

Gunjan Hariani was born to parents, Dhanraj Hariani and Sunita Hariani and has lived in Bombay (now Mumbai), India for the first 17 years of her life. After she finished her schooling in Bombay, she came to the cold but beautiful city of Houghton, MI to pursue her undergraduate degree in Bioinformatics at Michigan Technological University. She enjoyed her Bioinformatics curriculum at MTU and was very much blown away by the cool applications of science to answer various biological questions. Not surprisingly, she decided to pursue an advanced degree in Bioinformatics. While she was applying to graduate schools after her undergraduate degree, she interned at StoraEnso in yet another cold city of Wisconsin Rapids, WI where she enjoyed learning the process of making paper from pulp (she recommends the readers not to scratch their heads over the relationship between Bioinformatics and making paper from pulp). Well actually, during her internship she built her computational skills and setup a database to archive documents along with a web server to securely retrieve those documents. After her internship, she started her graduate degree in Bioinformatics at North Carolina State University in a not-so-cold city of Raleigh where she has learnt a lot academically and grown professionally with the guidance of her mentor, Dr. Alison Motsinger-Reif. Apart from learning, she likes to cook, listen to music, watch movies, and hang out with her friends in her spare time. She is a simple person and is easily amused with small things in life (like referring to herself in third person while writing her own biography). None the less, she has come to enjoy living in Raleigh for all the wonderful, inspiring, kind, and smart people she has met here.

ACKNOWLEDGMENTS

Foremost, I would like to thank Dr. Alison Motsinger-Reif for giving me an opportunity to work with her. I absolutely find it a privilege to have worked with an investigator who shows such jaw-dropping dedication towards her work – it is very inspiring. I would also like to thank her for being so patient with me while working on the sequencing dataset and I appreciate all the support that she has provided me over the years.

I also thank Drs. Zhao-Bang Zeng and Alison Motsinger-Reif for the funding support during my entire course of PhD. I am grateful to Dahlia Nielsen for the guidance she provided during the first years of my PhD and helped me land with Alison. I thank the rest of my committee members - Dr. Stone, Dr. Jorge and Dahlia for being so encouraging – it has meant a lot to me. I also appreciate them for their open door policy; and allowing me to interrupt them with questions without having to make any prior appointments.

Next, I would like to thank my colleagues at Expression Analysis, Inc for the knowledge transfer and all their help in accelerating my understanding on processing sequence data. Very special thanks to my supervisor Victor Weigman as he has let me harass him with questions via unlimited gchats and emails.

I also would like to thank Siarra Dickey for all the administrative help with my paperwork and for teaching me all sorts of American idioms. It was always a pleasure to stop by her office for a chat. I thank all my friends at BRC, especially Noffisat Oki, Ronglin Che, Monnat ‘Ginger’ Pongpanich – it has been a pleasure to get to know you all and spend time with you all. I would also like to thank Nicholas Hardison and Zeke Harris for all the help they provided with computer hardware and programming questions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Next Generation Sequencing technologies and their application to pharmacogenomics

Gunjan D. Hariani

Bioinformatics Research Center, North Carolina State University, Raleigh NC 27695

In this introductory chapter, I give a brief overview of the role that next generation sequencing (NGS) has played in variant discovery. I describe how this dissertation was started as an application of NGS for discovery in pharmacogenomics and the work that has been put in to set up pipeline for variant calling from NGS data; how I have compared different statistical methods for association testing of low frequency variants that can be detected through sequencing; and how I applied some of these methods to refine a previously acquired association signal in a pregnancy related trait (small for gestational age).

Role of sequencing in variant discovery

A primary goal of research in human diseases is to discover DNA variations that contribute to common diseases and traits of medical importance, to better understand disease etiology and the underlying genetic architecture behind it, improve pharmaceutical targets, and make advancements in prevention and cure of diseases. Our genomes harbor variations in various forms like single nucleotide polymorphisms (SNPs - the most common form of variation), mutations, insertions/deletions (indels), copy number variation, and chromosomal abnormalities. These variations can affect protein function in a myriad of ways (e.g. by directly distorting protein structure or indirectly through altered regulatory elements) and partly account for trait variability (environment, epigenetics, and gene-environment interaction are other factors that can influence a trait). If a variable trait shows considerable heritability, researchers are interested in knowing what genetic factors influence the trait and how they influence it; which makes cataloging of genetic variants an important step towards discovery.

A lot of effort has been put to develop technologies (and analytical methods to go along with these technologies) for variant discovery either through genotyping chips or

sequencing. On the sequencing forefront, we are in a new sequencing era and current technologies are offering solutions to make genomic exploration easier; we can zoom out and get a comprehensive view of our genome. Today, the second and third generations of sequencing technology (or NGS) have outpaced the capillary electrophoresis based Sanger sequencing technique (first generation of sequencing) not just in sheer volume of data output, but also in terms of number of samples that can be multiplexed (i.e. sequenced simultaneously on the same machine), time, and cost; albeit Sanger sequencing is the gold standard in terms of sequencing accuracy.

Apart from fast and relatively cheap data generation, DNA sequencing has become very popular due to its potential to discover all forms of genomic variations like indels, inversions/duplications, gene deletions, gene fusions, and copy number variations. It can detect previously uncharacterized variants– variants that are either private to an individual or have low frequency (less than 5%) in the population. As a result, it is not surprising that this technology has emerged quickly with numerous applications in association studies; rare Mendelian disorders (exome sequencing identified mutations on both copies of N-Glycanase 1 in the only person documented to lack this gene product (Need et al., 2012)); clinical diagnostics (whole genome sequencing identified an oncogene fusion event in a patient with acute myeloid leukemia and this finding altered her treatment plan (Welch et al., 2011)); assembly of new genomes (the genome of the giant Panda has been assembled using only short reads (Li et al., 2010)); assessment of gene orthology between species; somatic mutation discovery in tumors; and detecting DNA methylation; amongst many others.

DNA sequencing has especially become very popular in candidate gene association studies since researchers can now detect low frequency variants and examine their role in trait variability. This dissertation was initiated along similar lines to investigate the role of rare and low frequency variants along with common variants in the pharmacogenomics of cytotoxicity in an *ex-vivo* model (treat cell culture with a drug).

Application of NGS for discovery in pharmacogenomics

Pharmacogenomics is the study of how genetic variation in individuals affects their response to pharmaceutical drugs. Individuals show variable response to drugs ranging from no drug effect to adverse drug effects. Given this inter-individual variability, one of the goals in pharmacogenomics is to move towards personalized medicine and make informed decisions for prescription and dosing of drugs (Crews et al., 2012; Jonas & McLeod, 2009). This emphasizes the undertaking of such a study to associate genetic variants (common and rare) with drug response.

Drug response is a complex trait, i.e., multiple genes work together to manifest the trait. This is not surprising as a drug given to a patient has to be absorbed by the cells, distributed to tissues where it will be metabolized and then excreted from the system. As multiple genes are involved in these processes, there are numerous means by which the system can be broken. The result of a broken system may be lowered drug efficacy, ultra-fast drug metabolism or adverse side reactions; and hence the observed variability in response. With this understanding, the project was designed to look at 103 candidate genes involved in pathways for drug metabolism, transport, or drug action across 5 classes of chemotherapy drugs: fluoropyrimidines, anthracyclines, platinum compounds, taxanes, and camptothecins. Thanks to yet another evolved DNA amplification technology (RainDance technologies - RDT); we could target, capture and amplify all the 103 candidate genes within every cell line simultaneously thereby reducing the time/labor intensive part of the library preparation.

There were 30 FDA approved chemotherapeutic drugs chosen in this study. An earlier study established that most of these 30 drugs has high heritability (i.e., how much variation in a trait is due to variation in genetic factors) estimates (Peters et al., 2011). This justifies their use in this project as it only makes sense to look at genetic diversity if genetics is playing a role in trait variation; although this does not guarantee the ease to map genetic factors to the trait due to our complex genomes.

We were interested in cytotoxicity as drug response because chemotherapeutics are designed to kill the rapidly growing cells of the body. Hence, the drug response was measured as percent cell viability after treatment with the drug. Ninety-five lymphoblastoid

cell lines (LCLs) from 14 CEPH pedigrees were chosen for sequencing of candidate genes based on their *extreme* drug response (details are provided in Chapter 2). The sequencing was done on Illumina Genome Analyzer IIx (GAIIx) at four different centers with one sample sequenced per lane (no multiplexing was done).

Why this interest in uncommon variants and what statistical challenges arise when testing their association with phenotypes?

Chapter 1 gives a brief overview of the shift of paradigm from common variants to uncommon variants, and why researchers are interested in testing the role of low frequency variants with common complex traits. This interest has led to rapid development of statistical methods for association testing with rare variants. When the work on this chapter was started, to our knowledge only 3 statistical methods (CAST, CMC, and DWS – described in the chapter) were available and they were all designed for case-control phenotypes. I proposed an extension for the DWS test to quantitative traits (QWS) and compared its power to CMC and DWS using simulation studies.

I have incorporated the population genetics model proposed in (Pritchard, 2001) in simulating the allele frequency of the susceptible allele. Given the parameters used in his model, Pritchard suggests that a susceptible locus (gene) which contributes considerably to trait variability will display higher allelic heterogeneity than random loci and frequency of these polymorphisms will range from rare alleles (1%) to common alleles (50%). This theory advocates the role of rare variants in common diseases.

While the models simulated in chapter1 are simplistic and do not truly reflect the complex architecture of a disease model, these simulations were conducted in order to gain an idea on what sample sizes and effect sizes would be needed even under naïve assumptions for these statistical methods to show power for association. I conclude that these tests are suited for candidate gene studies but show considerably low powers for a GWAS setting. I also infer that unless the variants are showing strong effect in the same direction, small sample sizes will not be well powered for association testing with these methods.

Does it make sense to study uncommon variants along with common variants in pharmacogenomics?

This pharmacogenomics study is an exploratory initiative and the goal is to use this promising technology and improve our understanding of drug response through discovery of various variant forms. We now find evidence in recent literature about the role of rare variants in drug response (Avula R., Rand A., Black J.L., & O'Kane D.J., 2011; Fellay et al., 2010);which illustrates that a lot of work still lies ahead to put together the workings of these therapeutics. Research in this field has already led to changes in marketing of drugs and now an estimated 10% of FDA approved drugs have genetic testing labels (Frueh et al., 2008); this is exactly what personalized medicine is about.

In Chapter 2, I give a detailed description of the study design including the choice of *ex-vivo* model, how the 95 LCLs were selected, and how the drug response in the LCLs was measured. I also describe the pipeline that was set up to analyze the sequencing data from these 95 cell lines. A lot of time and effort has been put into setting up this pipeline and making sure that high quality variants are called. The pipeline is specific to processing Illumina based sequencing data. I have integrated our knowledge about characteristics of Illumina data and used this to build the pipeline – this includes quality score recalibration, low quality base trimming from the ends of the reads, and adapter clipping. Efforts were put in to assess quality at every step. Variants were called per sample; but I integrated variant quality statistics, variant calling information across samples, primer capture information, and family structure when filtering for high quality variants. I evaluated variant quality by incorporating information from different databases like dbSNP and Hapmap and assessing handful of variants in Integrative Genomics Viewer (Thorvaldsdottir, Robinson, & Mesirov, 2012). I conducted association analysis with Family Based Association Test (Laird, Horvath, & Xu, 2000) software and found significant hits for 2 of the 30 drugs used (oxaliplatin and bleomycin) after accounting for multiple hypothesis testing. These results are very exciting and show that an *ex-vivo* model can work well as a discovery tool in pharmacogenomics.

Can rare variants explain a Genome Wide Association Signal?

One of the promises of sequencing is that it can help with fine mapping of association signals robustly identified by Genome Wide Association Studies (GWAS) and also improve the understanding of the allelic architecture of variants in candidate genes associated with a trait. The signal can be refined by evaluating association and functional loss with variants that were untyped in the original GWAS but co-located with the associated marker. This approach has been applied to phenotypes like low-density lipoprotein cholesterol (Sanna et al., 2011), fetal hemoglobin levels (Galarneau et al., 2010), schizophrenia (Dow et al., 2011), type 2 diabetes (Shea et al., 2011).

We have used this concept to examine the role of variants in a candidate gene (paternally expressed 3 – PEG3) for a pregnancy related disorder - small for gestational age (SGA). The candidate gene was identified through a series of transcriptional gene expression and candidate SNP genotyping studies. We wanted to test if we could refine this previously identified association signal in PEG3 from 55 samples (23 Caucasians and 32 African Americans).This study was designed by Tsai et al(Tsai et al., 2010).

In Chapter 3, I give a comprehensive description of how we modified the variant calling pipeline from Chapter 2 and used it to call variants in the sequencing data. I annotated the variants and included this information in association testing of rare variants with SGA using one of the statistical methods tested in Chapter 1. While I did not find any significant accumulation of functional variants in SGA vs. non-SGA in either ethnic group; but found a significant excess of intronic variants in controls vs. cases for the Caucasian group. I conclude that this finding is a little difficult to interpret given that we had a small sample size (n=23). Apart from sample size restriction, difficult interpretation arises from the fact that the trait of interest is complex, difficult to characterize, and also the workings of an imprinted gene are complicated.

Conclusion

I end the dissertation by summarizing the findings of this research and commenting on some of the challenges that come with NGS data.

References

Avula R., Rand A., Black J.L., & O'Kane D.J. (2011). Simultaneous genotyping of multiple polymorphisms in human serotonin transporter gene and detection of novel allelic variants. *Translational Psychiatry, 1*(e32)

Crews, K. R., Gaedigk, A., Dunnenberger, H. M., Klein, T. E., Shen, D. D., Callaghan, J. T., et al. (2012). Clinical pharmacogenetics implementation consortium (CPIC) guidelines for codeine therapy in the context of cytochrome P450 2D6 (CYP2D6) genotype. *Clinical Pharmacology and Therapeutics, 91*(2), 321-326. doi:10.1038/clpt.2011.287; 10.1038/clpt.2011.287

Dow, D. J., Huxley-Jones, J., Hall, J. M., Francks, C., Maycox, P. R., Kew, J. N., et al. (2011). ADAMTSL3 as a candidate gene for schizophrenia: Gene sequencing and ultra-high density association analysis by imputation. *Schizophrenia Research, 127*(1-3), 28-34. doi:10.1016/j.schres.2010.12.009

Fellay, J., Thompson, A. J., Ge, D., Gumbs, C. E., Urban, T. J., Shianna, K. V., et al. (2010). ITPA gene variants protect against anaemia in patients treated for chronic hepatitis C. *Nature, 464*(7287), 405-408. doi:10.1038/nature08825

Frueh, F. W., Amur, S., Mummaneni, P., Epstein, R. S., Aubert, R. E., DeLuca, T. M., et al. (2008). Pharmacogenomic biomarker information in drug labels approved by the united states food and drug administration: Prevalence of related drug use. *Pharmacotherapy, 28*(8), 992-998. doi:10.1592/phco.28.8.992

Galarneau, G., Palmer, C. D., Sankaran, V. G., Orkin, S. H., Hirschhorn, J. N., & Lettre, G. (2010). Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nature Genetics, 42*(12), 1049-1051. doi:10.1038/ng.707

Jonas, D. E., & McLeod, H. L. (2009). Genetic and clinical factors relating to warfarin dosing. *Trends in Pharmacological Sciences, 30*(7), 375-386. doi:10.1016/j.tips.2009.05.001

Laird, N. M., Horvath, S., & Xu, X. (2000). Implementing a unified approach to family-based tests of association. *Genetic Epidemiology, 19 Suppl 1*, S36-42. doi:2-M

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., et al. (2010). The sequence and de novo assembly of the giant panda genome. *Nature, 463*(7279), 311-317. doi:10.1038/nature08696

Need, A. C., Shashi, V., Hitomi, Y., Schoch, K., Shianna, K. V., McDonald, M. T., et al. (2012). Clinical application of exome sequencing in undiagnosed genetic conditions. *Journal of Medical Genetics, 49*(6), 353-361. doi:10.1136/jmedgenet-2012-100819

Peters, E. J., Motsinger-Reif, A., Havener, T. M., Everitt, L., Hardison, N. E., Watson, V. G., et al. (2011). Pharmacogenomic characterization of US FDA-approved cytotoxic drugs. *Pharmacogenomics, 12*(10), 1407-1415. doi:10.2217/pgs.11.92

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *American Journal of Human Genetics, 69*(1), 124-137. doi:10.1086/321272

Sanna, S., Li, B., Mulas, A., Sidore, C., Kang, H. M., Jackson, A. U., et al. (2011). Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. *PLoS Genetics, 7*(7), e1002198. doi:10.1371/journal.pgen.1002198

Shea, J., Agarwala, V., Philippakis, A. A., Maguire, J., Banks, E., Depristo, M., et al. (2011). Comparing strategies to fine-map the association of common SNPs at chromosome 9p21 with type 2 diabetes and myocardial infarction. *Nature Genetics, 43*(8), 801-805. doi:10.1038/ng.871; 10.1038/ng.871

Thorvaldsdottir, H., Robinson, J. T., & Mesirov, J. P. (2012). Integrative genomics viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics,* doi:10.1093/bib/bbs017

Tsai, S., Hardison, N., Keebler, J., James, A., Motsinger-Reif, A., Bischoff, S., et al. (2010). Targeted solution hybrid capture, barcoding, and multiplexed massively parallel sequencing of PEG3 0.5 mb genomic interval in 55 human individuals. (PhD, North Carolina State University).

Welch, J. S., Westervelt, P., Ding, L., Larson, D. E., Klco, J. M., Kulkarni, S., et al. (2011). Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. *JAMA : The Journal of the American Medical Association, 305*(15), 1577-1584. doi:10.1001/jama.2011.497

CHAPTER 2

Power comparison of methods for associating rare variants with continuous traits on simulated models

Gunjan D. Hariani[1], Alison A. Motsinger-Reif[1,2]

[1]Bioinformatics Research Center, North Carolina State University, Raleigh NC, USA;

[2]Department of Statistics, North Carolina State University, Raleigh NC, USA

Abstract

With recent advances in genotyping technology, many genetic studies are investigating the role of rare variants in complex traits via the Common Disease Rare Variant (CDRV) hypothesis. In this study, I modify the existing weighted sums test designed for dichotomous traits (DWS) to test for excess of variants in the tails of quantitative traits and call this approach the quantitative weighted sums test (QWS). Powers of this test are compared with existing approaches like Combined Multivariate Collapsing Test (CMC) and weighted sums test on a range of simulated models with varying parameters. The results show that the powers of the DWS and QWS tests are comparable but better than CMC in these simulated models. A reduction in power is observed for all the tests on small cohort sizes. The power of these tests can be improved for small cohorts by designing prior hypothesis instead of conducting genome wide association analysis. The simulations also suggest that the pooling of variants by some biological function improves the power of the test and decreases spurious association with random non-causal variants.

Introduction

As it has been mentioned in the introduction of this dissertation, cataloging genomic variants is very valuable in genetics in order to conduct any trait association. Systematic approaches like the International HapMap Consortium that catalogs common genetic variations across populations have made it possible to successfully implicate SNPs with diseases via genome wide association studies (GWAS). Over the last few years, GWAS have been successfully used to detect genetic associations with a number on complex diseases,

including age-related macular degeneration (Klein et al., 2005), Type 1 diabetes (Todd et al., 2007), Type 2 diabetes (Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research et al., 2007) , prostate cancer (Gudmundsson et al., 2007), inflammatory bowel disease (Duerr et al., 2006) and more (Manolio, Brooks, & Collins, 2008).

While GWAS have had its share of successes, there are several limitations to this approach that have recently been highlighted in the literature (Frazer, Murray, Schork, & Topol, 2009). First, most of the variations discovered by this approach have small effect sizes for diseases and explain only a small proportion of estimated disease heritability (Manolio et al., 2009) for the diseases being studied. Additionally, the largest underlying assumption that GWAS rely on, the common disease-common variant hypothesis (CDCV) hypothesis (Reich & Lander, 2001) is likely not to be valid for many complex diseases and traits (Bodmer & Bonilla, 2008; Gorlov, Gorlova, Sunyaev, Spitz, & Amos, 2008; Pritchard & Cox, 2002). For example, it is speculated that the genetic architecture underlying schizophrenia consists of common and rare variants with small effect sizes instead of few prevalent risk alleles (Wray & Visscher, 2010).

In response to these limitations, and aided by recent advances in genotyping technology, many genetic studies are shifting gears towards investigating the common disease-rare variant (CDRV) hypothesis to help elucidate the unexplained variance of complex traits (Manolio et al., 2009). As per the CDRV hypothesis, no one common allele is responsible for disease susceptibility; rather multiple rare variants with moderate to high effect sizes result in diseased or extreme phenotypic outcome. Allelic heterogeneity plays an important role the CDRV dogma. Theoretical models (Pritchard, 2001) and increasing scientific evidence support this hypothesis (Cohen et al., 2004; Fearnhead et al., 2004; Ji et al., 2008; Nejentsev, Walker, Riches, Egholm, & Todd, 2009; Romeo et al., 2007). If the CDRV hypothesis indeed holds, it also helps explain the low power of GWAS and linkage mapping to find rare variants with modest effect sizes. Both these approaches work poorly in presence of allelic heterogeneity (Pritchard & Cox, 2002).

Recent advances in sequencing technologies have opened opportunities to investigate sequence variations at individual level and reveal rare and uncommon variants previously not captured by GWAS chips; e.g. 1000 genomes project (Siva, 2008). The association of such rare variants with traits and diseases of interest presents an important statistical and methodological challenge. While advances in technology have made sequence data more readily available than ever before, there are still practical limitations for very large sample sizes.

A few initial methodologies have been proposed for sequencing data applications for both binary and quantitative traits. The Combined Multivariate and Collapsing (CMC) method of Li and Leal involves collapsing the rare variants as per different frequency thresholds into groups and applying a multivariate test to check if any variants are associated with disease risk (Li & Leal, 2008). The Weighted Sums method proposed by Madsen and Browning detects excess rare variants in cases vs. controls by weighting the variants heavily if they are infrequently observed in the controls (Madsen & Browning, 2009).

In this study, I propose an approach for quantitative traits by extending the Weighted Sums approach (Madsen & Browning, 2009) and demonstrate its performance by evaluating both power and Type I data in a wide range of simulated data with varying genetic models. Additionally, I evaluate different collapsing techniques incorporating putative functional information about variants into the collapsing procedures. These different strategies are evaluated for both quantitative and threshold-based binary traits (by choosing extreme percentiles as cases). I show that dichotomizing the trait is not significantly different than the proposed test and yield similar results in terms of power. Also, we demonstrate that the collapsing scheme for rare variants plays an important role in improving the power of the test and controlling the Type I error for association.

Methods

In the current study I propose and evaluate two different weighting schemes extended from the Weighted Sums method (Madsen & Browning, 2009) to quantitative traits. Understanding that a variety of genetic etiologies could explain the etiology of complex

traits, the two different schemes were designed to capture association signals for different genetic hypotheses. I have tried to encompass a range of models (distributions with low or wide spread, variants causing shift in one or both directions of phenotype, etc.) and test which method works in each situation. The two tests implemented in this study include Quantitative Weighted Sum (QWS) and Folded Weighted Sum (FWS).

The idea behind these methods is to test if the presence of a rare susceptible allele has an effect on phenotype. They check whether enrichment of rare variants causes a shift (increase or decrease) in phenotype from the underlying distribution where no rare alleles are observed. Under the null hypothesis, there is independence (no correlation) between rare variants and phenotype. The two weighting schemes are described below:

Quantitative Weighted Sum Test (QWS):

QWS is designed to work for cases where the variants cause a one-way shift in phenotype, i.e., variants either increase the trait value or decrease it. The basic idea is to compute a weight for each variant and correlate the weights with individual phenotypes. Under the null hypothesis, the variants are spread evenly across the phenotype distribution instead of being enriched in the extremes of the distribution. As a result, the phenotype has no (or low) correlation with the variants. The following steps are involved in computing the correlation in QWS method:

1. As suggested in (Madsen & Browning, 2009), collapse the rare variants in an individual by a biological category (synonymous/non-synonymous change, intronic/exonic variations, amino acid properties, and so on) or putative function (gene, pathway, exons, etc). There are a few advantages to grouping the variants: Firstly, the number of hypotheses to be tested is greatly reduced when variants are grouped together and this increases the power to detect correlation of variants with phenotype. Secondly, the collapsing of variants by functions like synonymous/non-synonymous decreases spurious association of variants with phenotype. Given a certain pooling function, let us assume that there are $i = 1,2,\ldots,L$ variants in a group.

2. Divide the distribution into cases and controls: choose the people from tails of the distribution as cases and the other percentiles as controls. For example, if it is believed that the variants cause an increase in phenotype then select the 90^{th} percentile as cases. If the effect of rare variants on phenotype is not clearly known, then chose the 10^{th} and 90^{th} percentiles as cases. For every variant *i*, compute a weight (w_i) as described in (Madsen & Browning, 2009). The weight for a rare variant indicates how often that variant is observed in the tails vs. other percentiles of the distribution. If a variant is highly prevalent in the tails of the distribution when compared to other percentiles, then the variant gets a high weight. The same applies if a variant is observed more often in percentiles other than the extremes; making the weights symmetrical in nature. The sign of the correlation value can be used to determine the direction of shift by variants in the trait. Sensitivity analysis with different cut-offs (percentiles) was performed to see how the choice of cut-off percentile for cases affects the power of the method. The weights are used to compute correlation of rare variants with the phenotype.

3. Calculate a genetic score for every individual *j* depending upon the number of rare susceptible alleles carried by that person. The genetic score can be interpreted as the mutational load for carrying susceptible alleles and is calculated as follows:

$$G_j = \Sigma_i (T_{ij}/w_i)$$

Here, T_{ij} is the total number of susceptible alleles (0, 1 or 2) carried by an individual *j* at variant *i*.

4. Compute Spearman's rank correlation between the genetic score and phenotype for all individuals. The choice of Spearman's rank correlation seems appropriate because the genetic scores do not follow a normal distribution (most people with no rare variants end up with a genetic score of zero). The p-value associated with the Spearman's rank correlation test is used to reject the null hypothesis of independence between rare variants and the phenotype.

Folded Weighted Sum Test (FWS):

The FWS test is modified from QWS to work for variants that shift the phenotype in both directions, i.e., some variants increase the phenotype of an individual and others decrease the phenotype. FWS differs from QWS in Step 4 which is described below.

In step 4, the original distribution is folded before computing the correlation. The folded distribution is derived by taking the absolute difference of each trait value and the mean of the trait distribution. These new values are used to compute Spearman's rank correlation with the genetic scores of the individual. By folding the distribution, the direction of change for all variants is synchronized and the power to find correlation of variants with the phenotype is retained.

Simulations

To compare and evaluate the methods described above, simulations were performed for a range of genetic models, with different allele frequencies, number of susceptibility alleles, presence of neutral variation(s), and different cohort sizes. For each model, a total of 1000 replicates were used to assess power by computing the proportion of replicates that yield a significant association at α=0.05. Power across different tests were compared using analysis of variance (ANOVA) controlling for model type, number of neutral variants and cohort size and Tukey correction for multiple testing was applied. Type I error rate was determined by using 1000 replicates of a null model (with no risk variants) – model in which the phenotype of an individual is sampled independent of the variants carried by that individual, as a result of which the variants are spread randomly throughout the distribution. Type I error was computed as the proportion of statistically significant results in the 1000 replicates of null model. The parameters considered in the data simulation are summarized in Table 2.1.

Frequency spectrum of rare alleles

As described in (Pritchard, 2001) assume that every variant has two types of alleles – normal allele (*N*) and rare susceptible allele (*S*). The frequency of the *S* allele for each variant

is sampled from the stationary probability distribution, f(p) given by Wright's formula (Wright, 1949; Ewens, 1979) as enumerated in (Pritchard, 2001)

$$f(p) = kp^{(\beta S-1)}(1-p)^{(\beta N-1)}e^{\sigma(1-p)}$$

Here, k is the normalization constant; p is the population frequency of the rare allele. β_S, β_N and σ represent rates for forward mutation, backward mutation and selection respectively, each rescaled by a factor of $4N_e$ (Pritchard, 2001). The parameter values are chosen as β_S = 0.001, $\beta_N = \beta_S/3$ and $\sigma = 12$ to allow for weak purifying selection but yet maintain the rare alleles in the population (Madsen & Browning, 2009; Pritchard, 2001; Pritchard & Cox, 2002). From the sampled points, frequencies that fall in the range of 0.01 – 0.001 are called as rare and anything below 0.001 is very rare. Given the small cohort sizes (100, 250, 500, 1000) considered in our simulations, only the frequencies from rare category were used to enrich the cohorts with *S* alleles. The number of alleles and cohort sizes simulated are summarized in Figure 2.1.

Sampling of individuals and phenotypes

A population of 10,000 alleles was generated per variant, according to the frequency of the *S* allele for that variant. Two alleles were randomly brought together to create a genotype for a given variant locus. Each locus was sampled independently under the assumption of no LD because new mutations usually arise on different haplotypes (Pritchard, 2001).

For the simulation of phenotype for an individual, genotypes across all variants were collapsed and the total number of S alleles discovered was used to model the phenotype. When no susceptible alleles are detected, the phenotype is sampled from a normal distribution of mean μ_0 and standard deviation σ [$N(\mu_0, \sigma)$]. When *n* susceptible alleles are detected, the phenotype is sampled from $N(\mu_0+nx, \sigma)$; where *x* is the shift in phenotype that occurs due to the addition of a susceptible allele. As it can be seen, the underlying genetic model is assumed to be additive in our simulations. For sake of simplicity, each variant is

assumed to have the same effect on phenotype, i.e., the shift in phenotype (x) associated with each variant is constant.

Simulated Trait Models

Four different models were simulated by varying the sampling distribution of phenotypes. The parameters x and σ of the normal distribution were modified to create the different models. These parameters play an important role in determining whether the tails of the distribution will be enriched with rare variants or if the variants will be more spread out over the different percentiles. The simulated models are summarized in Figure 2.2.

Model 1 is simulated with $x=2$, $\sigma=1$; resulting in densely packed variants in the tail of the overall distribution. Contrastingly, Model 2 ($x=2$, $\sigma=2$) shows a considerable spread of the variants over the different percentiles of the distribution. Model 3 ($x=3$, $\sigma=2$) lies intermediate between Models 1 and 2; the rare variants are not as compactly packed in the tails like Model 1 but also do not have a spread as drastic as Model 2. Model 4 ($x=4$, $\sigma=3$) follows Model 3 but the overall distribution has a higher variance as compared to Model 3.

Method Implementation

The four tests are implemented in Perl v5.10.0 and SAS 9.1.3 Service pack 4. To run QWS on a 100 replicates of Model 1 with cohort size of 1000 and 20 variant loci, 14 seconds were needed on Ubuntu Linux 9.04 with dual memory processor (Intel(R) Xeno(R) E5450) and 32 GB of memory. The ANOVA was done in SAS 9.1.3 (www.sas.org).

Results

Power for different models

I compare power and type I error rates of the different tests implemented (QWS, FWS, and CMC) for the four models described in Simulations section. I also dichotomize the trait into cases (>= 90^{th} percentile) and controls (< 90^{th} percentile) and apply the weighted sums described in (Madsen & Browning, 2009) and call this approach the Discrete Weighted Sums test (DWS). For bimodal distributions, where the variants shift the phenotype in both

the directions, I classify the 10^{th} and 90^{th} percentiles of the trait distribution as cases. One sided p-values for DWS (to test the alternate hypothesis that cases have greater number of variants than controls) are used in power comparison with QWS and FWS. For CMC power calculations in bimodal distributions, the folded phenotype is used to conduct t-test after collapsing at a specified threshold. Power for each different method was calculated and compared at two different alpha levels – 1) nominal level for a single gene (0.05) and 2) GWAS levels assuming 20,000 genes and using Bonferroni correction for multiple tests (0.05/20000).

Results are organized in five parts: Section 1 focuses on distributions shifted in one-direction due to causal variants only; Section 2 concentrates on bi-directional change simulations based on causal variants only; and Section 3 looks at distributions shifted in one direction by including neutral variants along with causal ones. Sections 4 and 5 talk about Type I error rates and choice of cut-off percentiles to determine weights of the variants.

Section 1: Shift of phenotype by variants in one direction only

This section focuses on one-way phenotype changes based on variants that belong to the rare category (0.1-1%), moderately-rare frequency variants (1-5%) and a mixture of rare and moderately-rare variants – for every three rare variants, two variants belong to low frequency.

Causal rare frequency variants (0.1-1%): I present results for GWAS corrected levels as the methods (CMC, DWS, QWS) don't shown significant differences in performance at nominal levels. In general, as expected, power for all the methods improves with increase in cohort size and enrichment of variants (Figure 2.3). For model 1, CMC is significantly different from DWS (Tukey corrected pval: 0.0425). For models 2-4, QWS vs. DWS and CMC vs. QWS show no statistical difference in power for associating the trait with variants. FWS does not perform well in these simulations.

The spread of the variants itself has an impact on the power of the methods. The power of QWS and DWS drops dramatically when the spread of variants is very high as

compared to when the variants are concentrated in the tails of the distribution (Model 2 vs. Model 1). An ANOVA comparing Models 3 and 4 did not yield any significant differences between power of these two models, after accounting for methods, cohort size, number of causal variants simulated and interaction between method and model.

It doesn't come as a surprise that FWS does not perform as well as the other methods for variants acting in one direction – FWS folds the distribution before computing correlation which reduces the original signal of association in this scenario (See Methods).

Causal low frequency variants (1-5%): After accounting for other factors, power of DWS is significantly different from CMC and QWS. CMC, QWS show high power at GWAS levels for all models enriched with moderately-rare variants (Figure 2.4). DWS shows low power for a cohort size of 100 and the power decreases with increase in number of variant loci. However, power for DWS to find excess of variants in tails of distribution vs. other percentiles increases dramatically with enlarged cohorts.

Causal mix frequency variants (0.1-5%): The results for mix frequency show no significant differences in powers of QWS, CMC, and DWS after accounting for model, cohort, and number of variants (Figure2.5).

Section 2: Shift of phenotype by variants in either direction of the distribution

This section only discusses the scenario for rare frequency variants (0.1-1%). Two cases are considered – In the first case, 50% of the variants shift the phenotype to the right and 50% shift it to the left (50-50) (Figure 2.6). In the second scenario, 75% of the variants shift the phenotype to the right and 25% to the left (75-25) (Figure 2.7). For the 50-50 models, CMC is significantly different from FWS (Tukey pval: 0.0010). QWS has no power in these simulations. For the 75-25 models, CMC is significantly different from all the other three tests; DWS is not significantly different from FWS; and QWS has no power to detect association. When comparing the 50-50 models to 75-25 models, only QWS shows significantly different patterns of performance ($p<0.001$) (QWS has higher power in 75-25 relative to 50-50 distribution). QWS shows differences in power because when most of the

variants are concentrated on one side of the distribution as in 75-25, QWS can get relatively high correlation values as opposed to when variants are equally concentrated in both the tails of the distribution.

For the bimodal simulations, CMC shows better power than DWS and FWS. Other observations for bimodal distribution with rare variants are very similar to rare variants causing a one-directional shift in phenotype. All the methods quickly lose power when the variants have a high spread and are not concentrated in the tails of the distribution (Model 2 vs. Model 1). Only QWS show differences in powers when comparing Model4 vs. Model3 where the overall distribution has a greater spread Model 4 as compared to Model 3.

Section 3: Addition of neutral variants

The addition of neutral variants decreases the power of the methods to detect association of variants with the phenotype along with increasing the Type I error rates of the methods. However, pooling neutral variants separately from causal variants helps restore the power of the methods (Figure 2.8).

Section 4: Type I error rates

Type I error rates for these tests were determined by simulating models with five neutral rare variants independent of the phenotype (Figure 2.9). Overall, it can be seen that type I error rates are well controlled for CMC, QWS and FWS across different models and frequency spectrum. DWS generally shows a slightly higher false positive rate.

Section 5: Choice of cut-off percentiles

The choice of cut-off percentiles does not affect the power for QWS but it does affect DWS at different percentiles (Cutoff percentiles of 10, 70, 80, and 90 were compared). The choice of 70^{th} percentile resulted in differences in power when compared to 90^{th} percentile for DWS. The choice of 10^{th} percentile resulted in no power for DWS. When I compare DWS and QWS at different percentiles, no significant differences in power is observed.

Discussion

The recent exploration of CDRV hypothesis has resulted in some rapid method developments in this area. In this study, I propose a Quantitative weighted Sums test and compare results with dichotomizing the traits. As expected, the power of these tests itself depends on a variety of factors – variance of overall phenotype, spread of the rare variants across different percentiles, the direction of spread by variants (uni-directional or bi-directional), presence of neutral variants, size of the cohort, and the number and frequency of variants, to name a few. I discuss the effects of these variables on power to detect association in the following paragraphs.

The power of the QWS drops if the variance of phenotype or the spread of the variants is very high. This is easily understood because the test is designed to find excess of variants in the tails of the phenotype distribution and the power of the test diminishes if the variants are not concentrated in the tails. The direction of change by variants also affects the power of QWS – the test does not work if the variants cause a change in both the directions of the phenotype. Folding the trait distribution and applying the CMC method works well in this case. But, one needs to know the frequency of variants at which you want to collapse the genotypes.

The size of the cohorts is important in deciding which tests will work well. For large cohort sizes, the QWS test shows more than 70% power to detect association of rare variants with the phenotype. Dividing the continuous traits into cases and controls does not perform any better than QWS. But, dichotomizing the trait comes at the cost of being able to recognize the appropriate cut-off percentile for cases. If the cut-off point for cases is not precise, then the power for the test is reduced. An alternative way of improving power for small sample sizes is to develop prior hypothesis and conduct candidate gene studies to avoid hypothesis testing at GWAS levels. By comparing powers from GWAS levels to nominal levels for all methods, it was observed that at nominal p-value, these methods show relatively higher power to detect association of variants with trait for small cohort sizes.

The addition of neutral variants increases the Type I error rates – this is also suggested in (Li & Leal, 2009). A decrease in power is also observed with accumulation of

neutral variants. However, it can be seen that if the neutral variants are pooled separately from causal variants, the power to detect association can be improved. Neutral variants could be pooled based on functional change of gene or protein with respect to the variant, frequency similarity of the variant in enhanced phenotype group and control phenotype group, synonymous or non-synonymous substitution and so on. Grouping variants on such criteria is important to avoid false association of a variant with the trait.

Unsurprisingly, the number and frequency of the causal variants also affects the power of the methods. The methods in general show an increase in power with increase in number of variants across rare frequency model simulations. DWS shows peculiar power patterns in presence of low frequency variants. Its power decreases with increase in variant counts for small cohort sizes; but the power dramatically increases for big cohorts. The inverse relationship between power of DWS and number of variants for a small cohort can be explained as follows – most of the percentiles of the distribution are enriched with variants; leaving no difference in variant frequency between extreme percentiles and other percentiles. For the big cohorts, an augmentation of variants across the distribution is observed, but DWS still shows excess number of variants in case group as compared to control group. This is because the ranked sums statistic very unstable with the combination of low frequency variants and an imbalance in cases/controls. Hence, the use of DWS test is not recommended in scenario where low frequency variants might shift the distribution. It may be noted that QWS does not utilize the ranked sum statistic and is not affected by low frequency variants.

In this study, I make a simple extension of the Madsen and Browning method to quantitative traits and evaluate its performance under varying parameters. I conclude that the success of any of these tests highly depends upon the underlying genetic architecture – if the rare variants really shift the phenotype to the extreme levels then these tests will work well; otherwise they will fail to detect true association. This emphasizes the importance of research effort to better understand the mechanism of trait etiology the rare variants for a given trait and develop methods suitable to capture signal for that system. Elucidation on the workings of these variants will play a key role in designing apt statistical methods that associate variants with traits.

Table 2.1: Description of different parameters used to simulate the models. All parameters considered in the current study are listed in this table.

Parameter	**Parameter Definition**	**Parameter Values**	**Additional Comments**
Cohort Size	Total number of people sequenced for the phenotype of interest	100, 250, 500, 1000	
Causal variants observed	Based on the collapsing technique (gene, pathway, etc.), the number of causal variants observed in the group	5, 10, 15, 20	
Neutral variants observed	Based on the collapsing technique (gene, pathway, etc.), the number of neutral variants observed in the group	5	
Frequency of the susceptible allele (*S*)	Frequency of the *S* allele was categorized as rare (0.1-1%), moderately-rare (1-5%), a mix of rare and moderately-rare (<5%)	0.1-1%, 1-5%, <5%	For the mixed frequency, the ratio of rare to moderately-rare alleles is 3:2
Direction of change	Two scenarios are considered: 1. The variants shift the phenotype in one direction only 2. The variants shift the phenotype in both directions	3:1 1:1	The parameter value applies only to the second scenario and indicates the ratio of variants that increase phenotype to variants that decrease phenotype

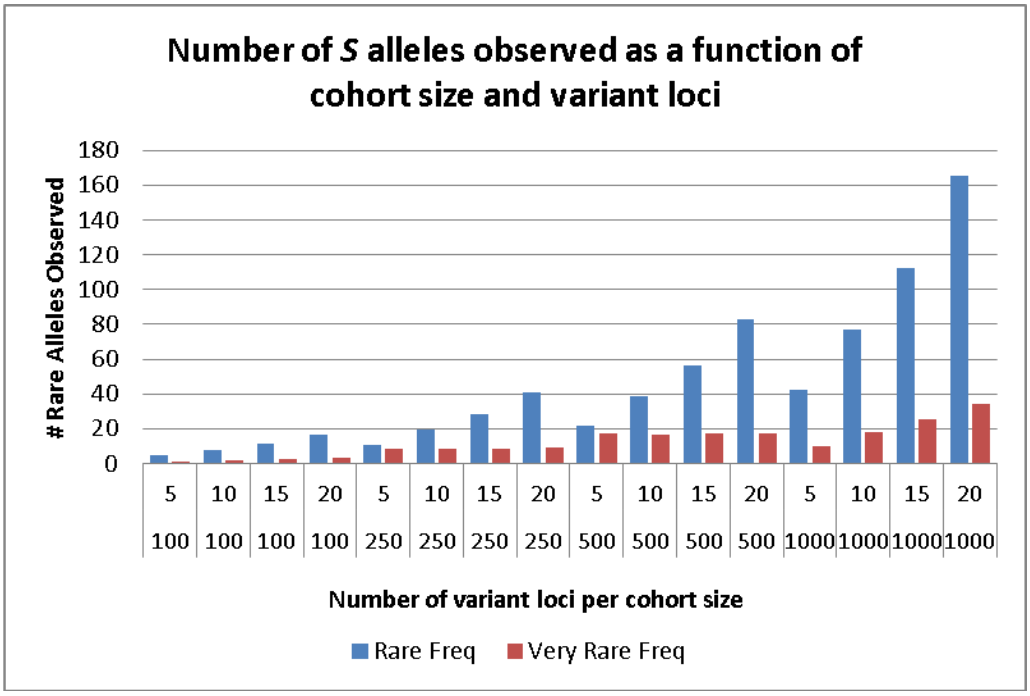

Figure 2.1: Number of rare susceptible alleles simulated. Number of rare susceptible alleles observed in a particular cohort dependent on the number of variant loci is shown. The x-axis shows the number of variants simulated (5, 10, 15, or 20) and the cohort size (100, 250, 500, or 1000). For the very rare frequency range (< 0.001) indicated in red, the cohorts are not enriched for variants. On an average from 100 replications, only 35 rare alleles were observed by sampling a thousand people with 20 variant sites when rare allele frequency is <0.001.

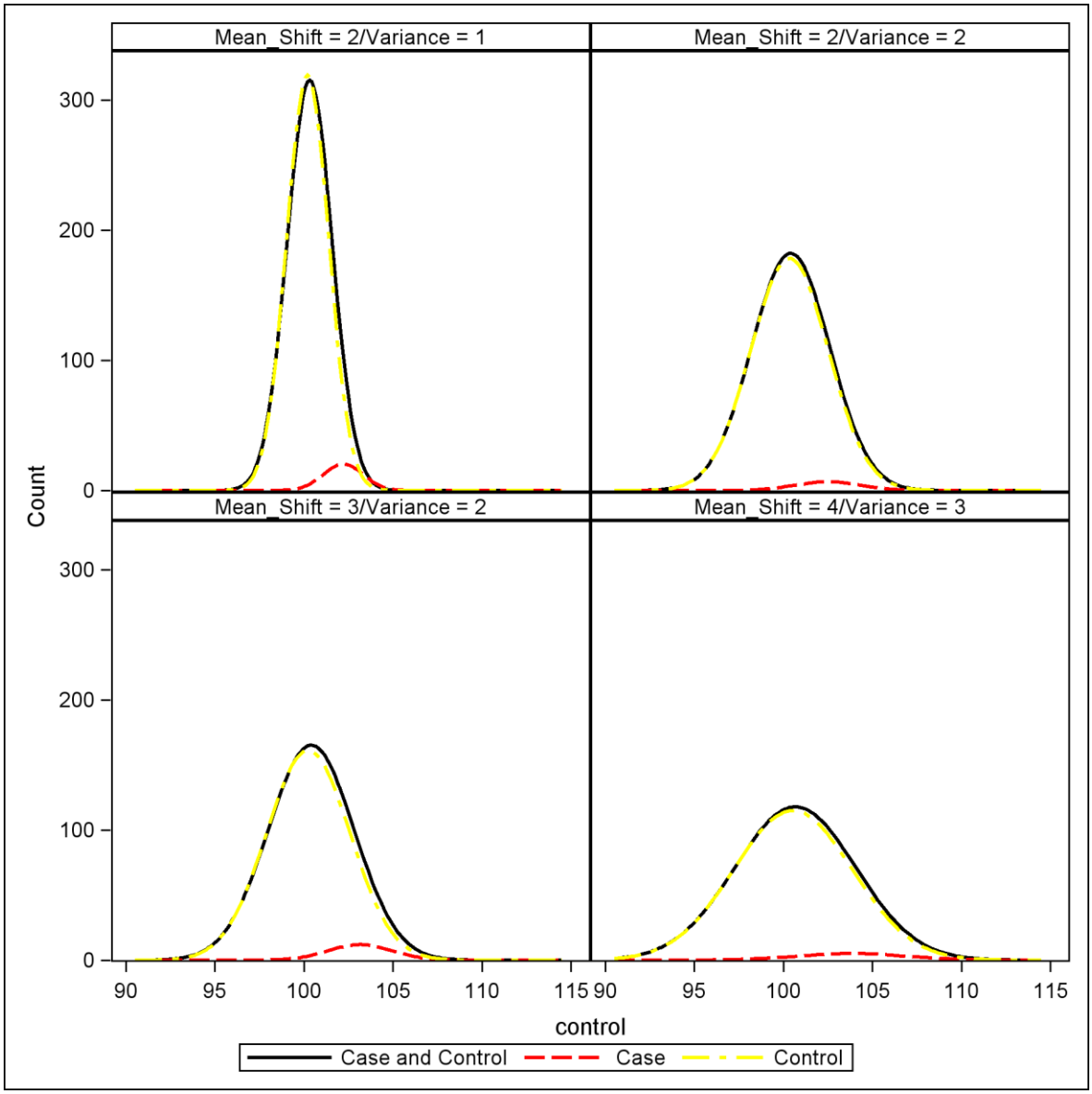

Figure 2.2: Different models simulated in this study. Each distribution (black) is created from sampling 1000 individuals with 20 variant sites. The red-dashed histogram represents the phenotypes of people with one or more rare variants.

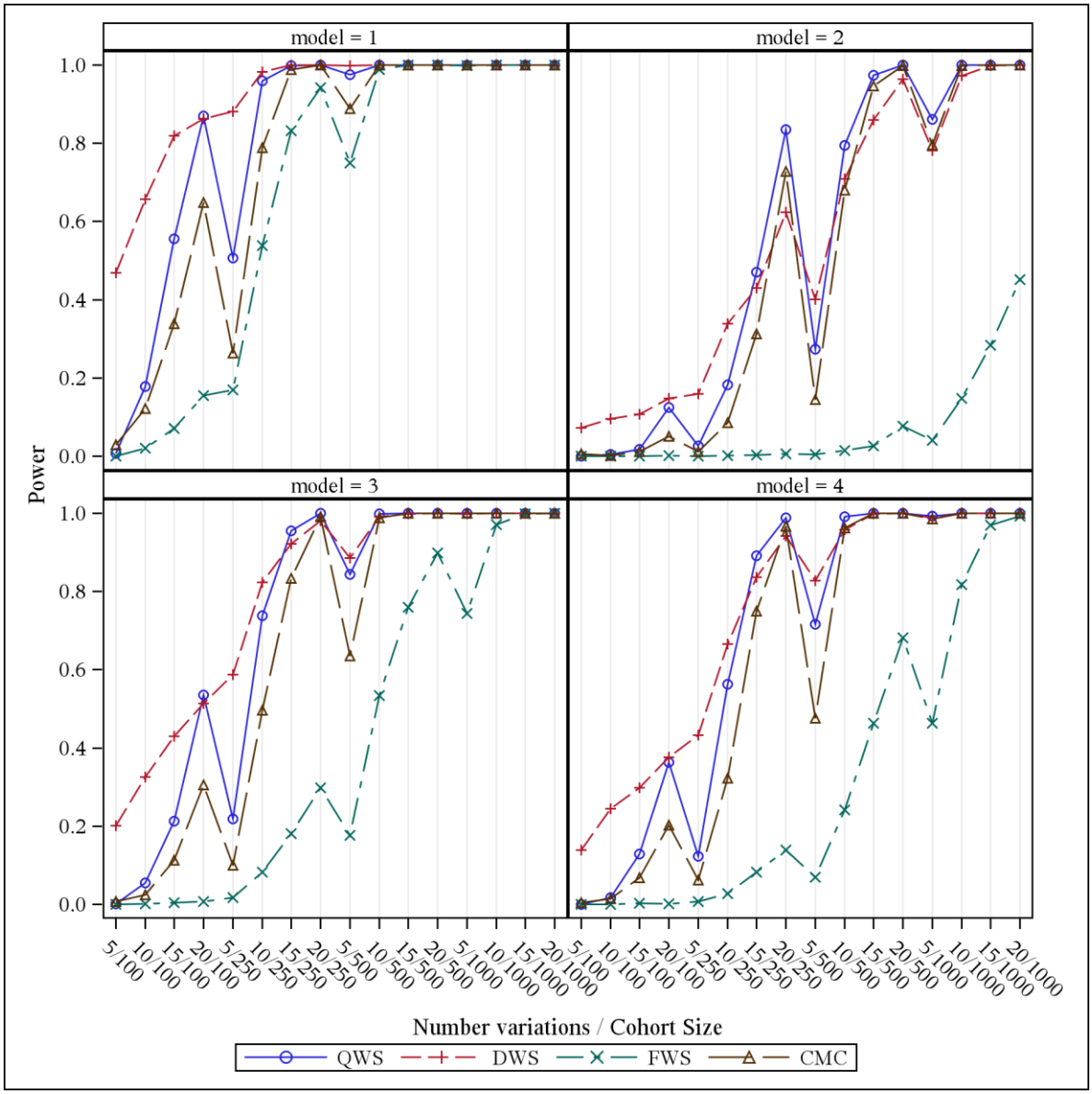

Figure 2.3: Power results for rare-frequency and one-direction shift. Powers at GWAS levels (0.05/20,000) for CMC, QWS, FWS and DWS for causal rare frequency variants causing a one-directional shift in phenotype are shown.

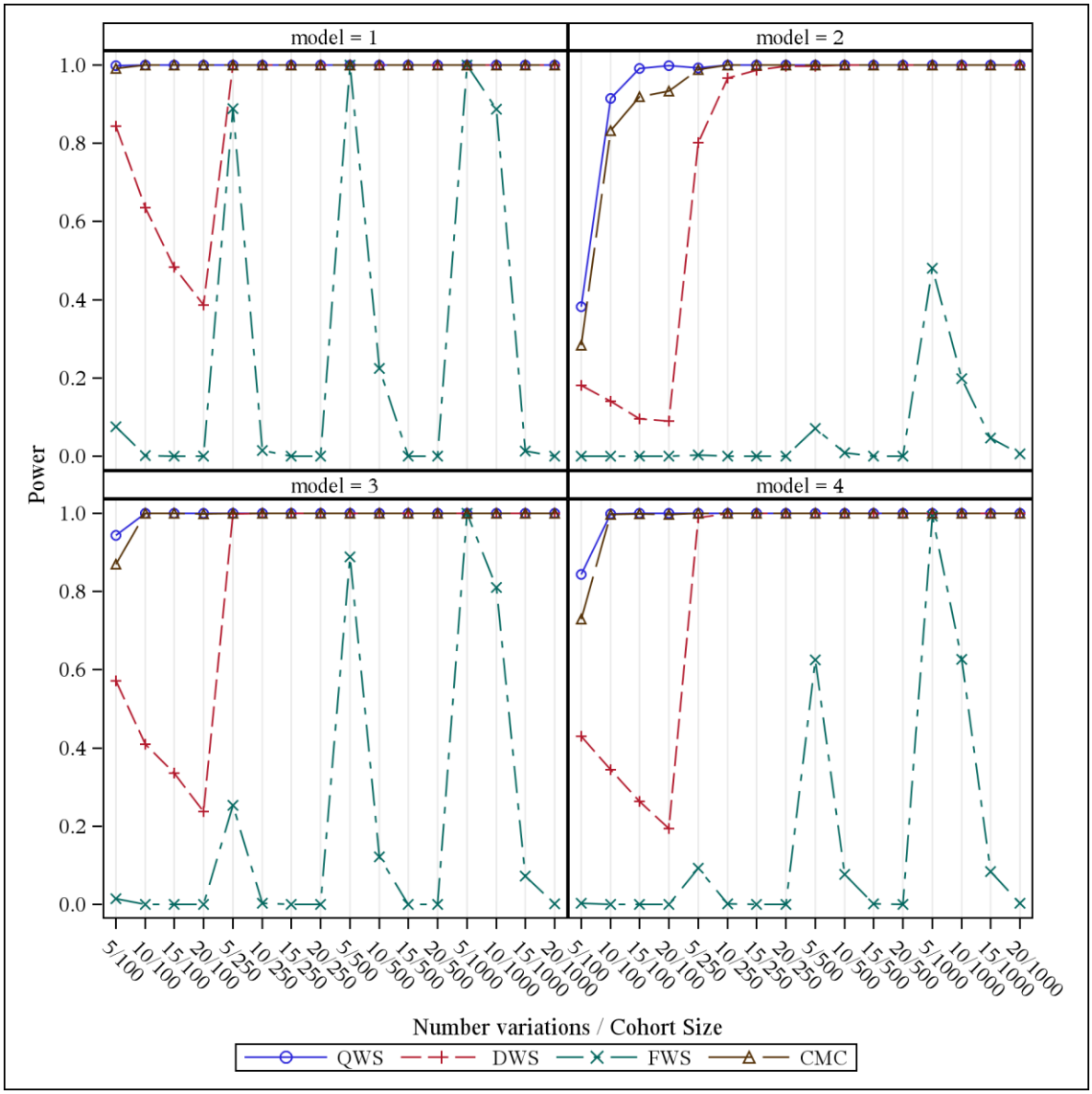

Figure 2.4: Power results for low-frequency and one-direction shift. Powers at GWAS levels for CMC, QWS, FWS and DWS for causal low frequency variants causing a one-directional shift in phenotype are shown.

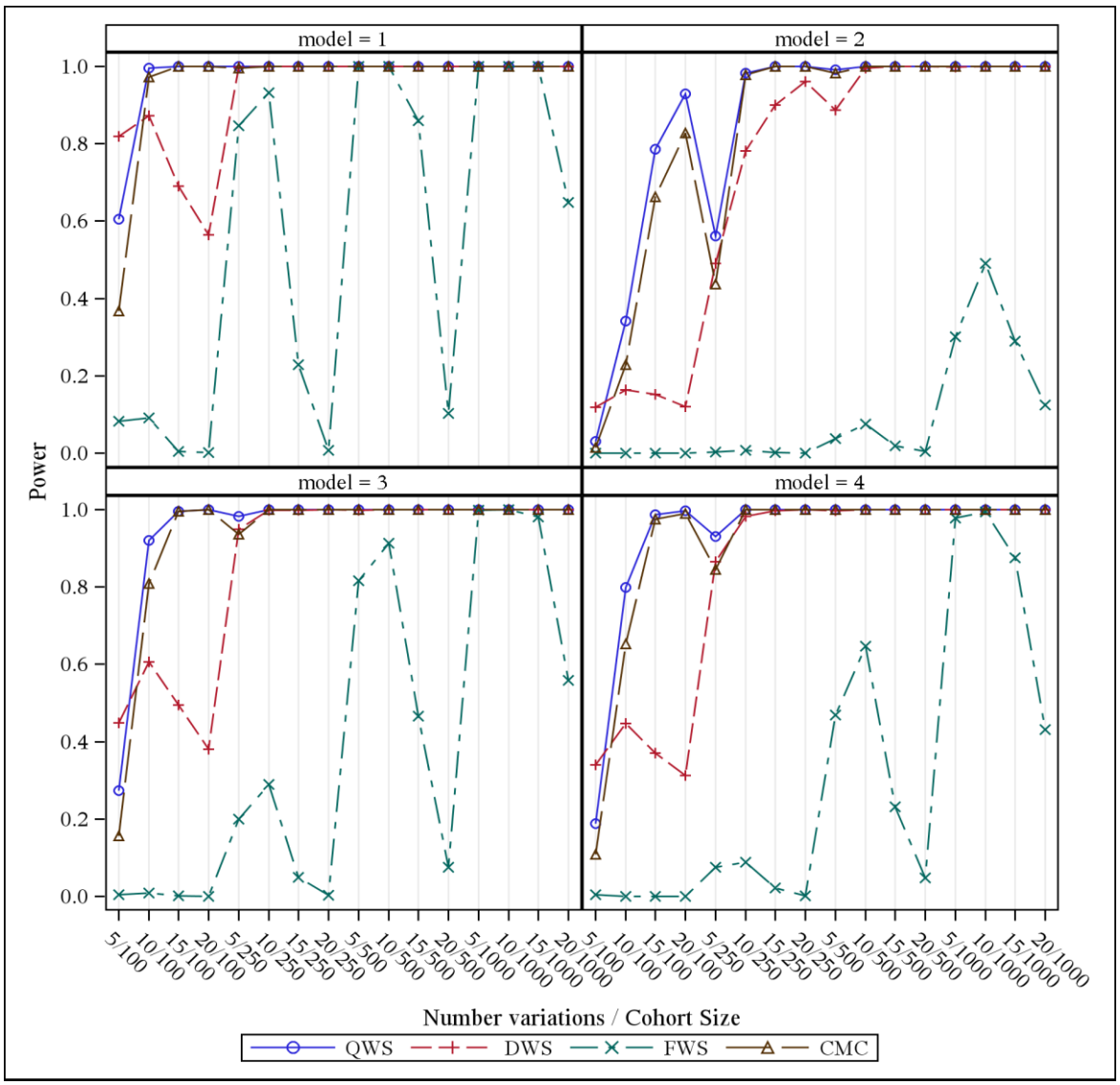

Figure 2.5: Power results for mix-frequency and one-direction shift. Powers at GWAS levels for CMC, QWS, FWS and DWS for causal mix frequency variants causing a one-directional shift in phenotype are shown.

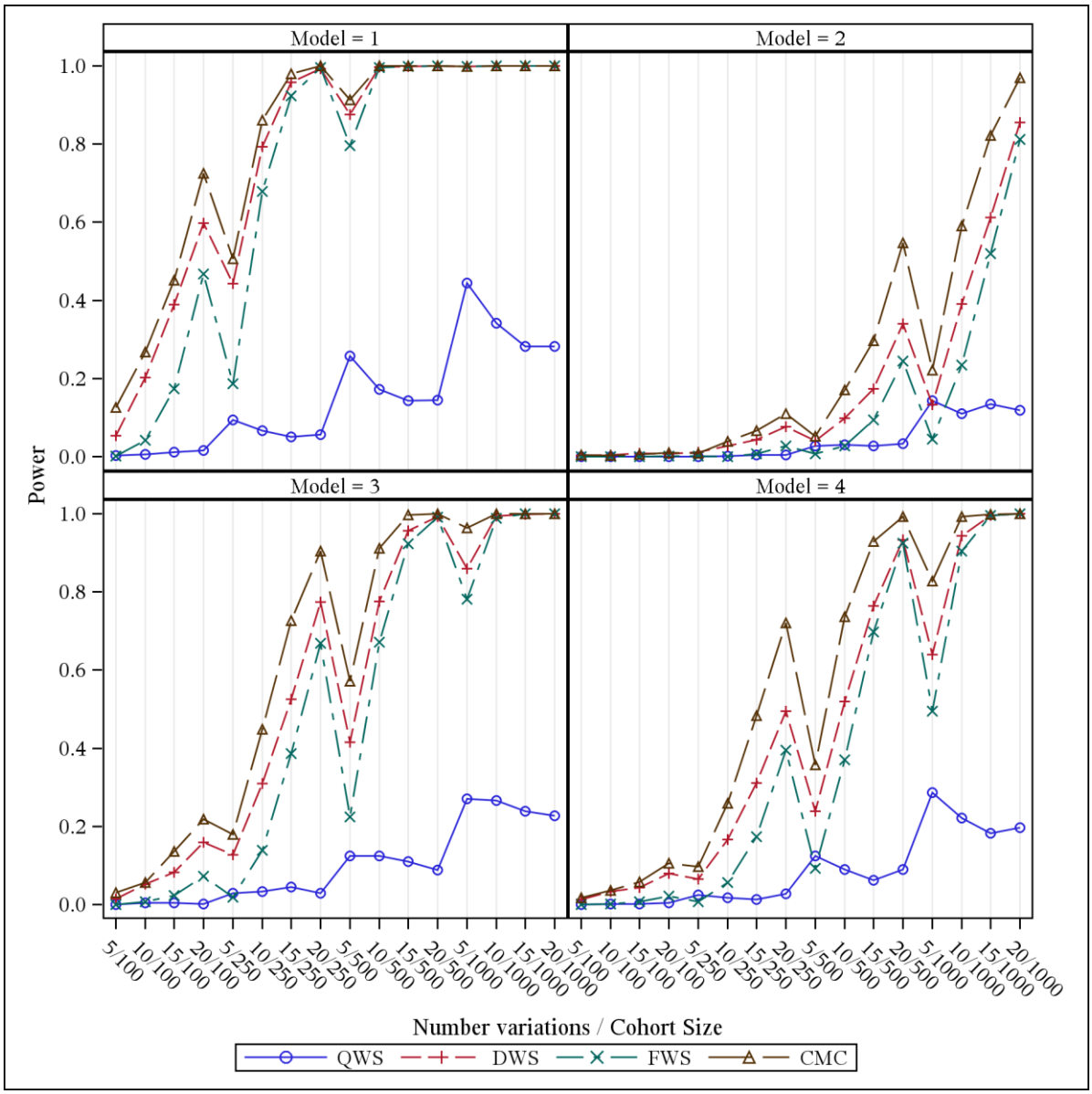

Figure 2.6: Power results for rare-frequency and bimodal-direction shift (50/50). Powers at GWAS levels for CMC, QWS, FWS and DWS for causal rare frequency variants causing a bi-directional shift in phenotype are shown. Fifty percent of the variants shift the phenotype to the right; while the remainder fifty percent shifts it to the left.

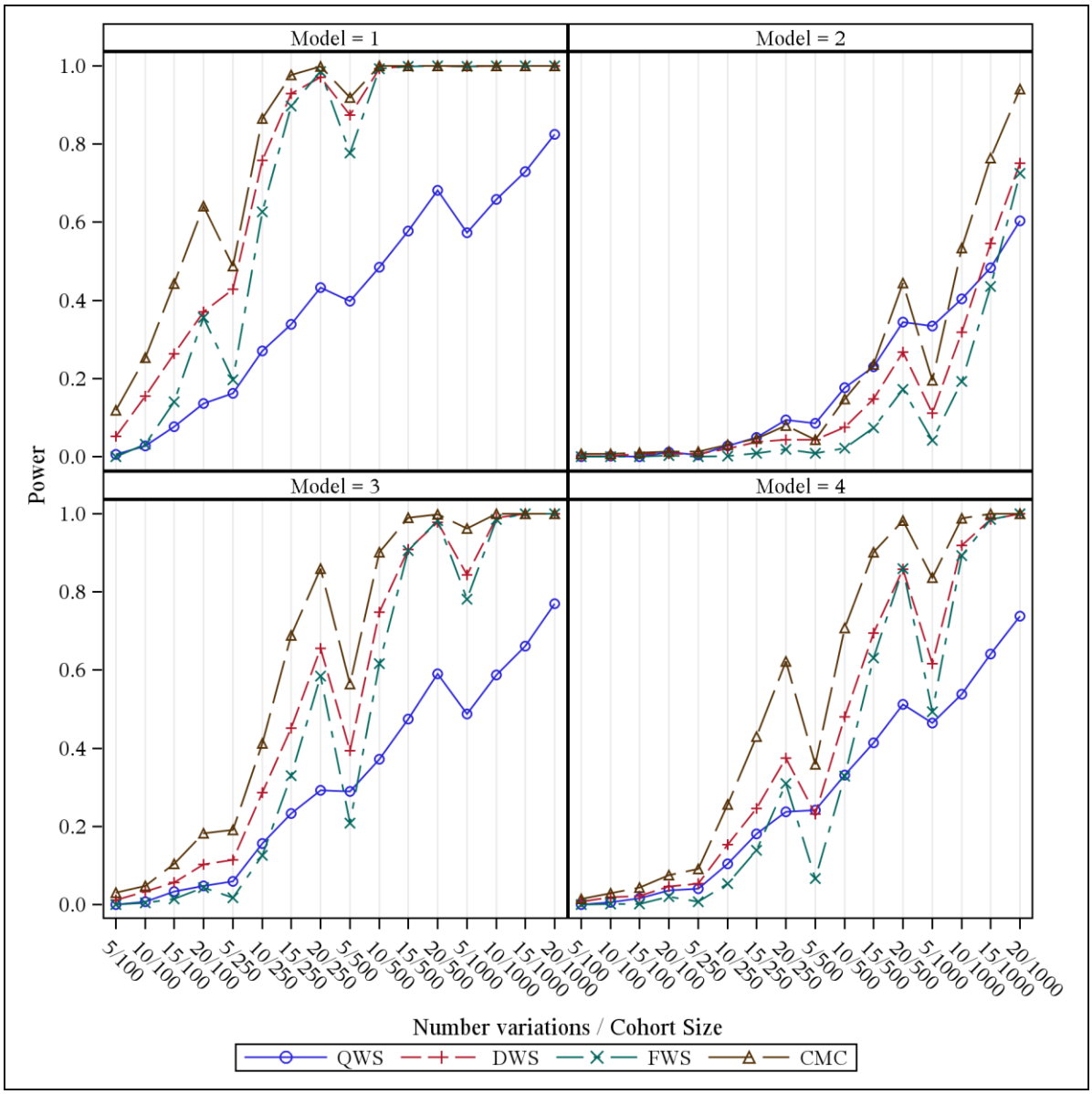

Figure 2.7: Power results for rare-frequency and bimodal-direction shift (75/25). Powers at GWAS levels for CMC, QWS, FWS and DWS for causal rare frequency variants causing a bi-directional shift in phenotype are shown. In this case, 75%of the variants shift the phenotype to the right and 25% of the variants shift it to the left.

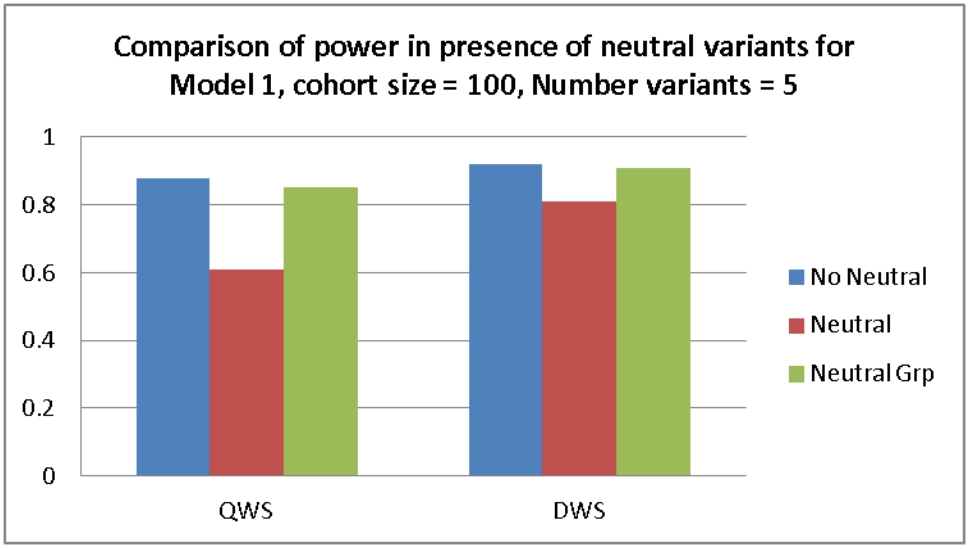

Figure 2.8: Power comparisons for simulations with neutral variants and no neutral variants. The figure shows a specific example from Model 1 with cohort size of 100 and 5 rare variants. Powers are presented at nominal levels. Adding of neutral variants decreases the power but grouping them separately from causal variants improves association efficiency.

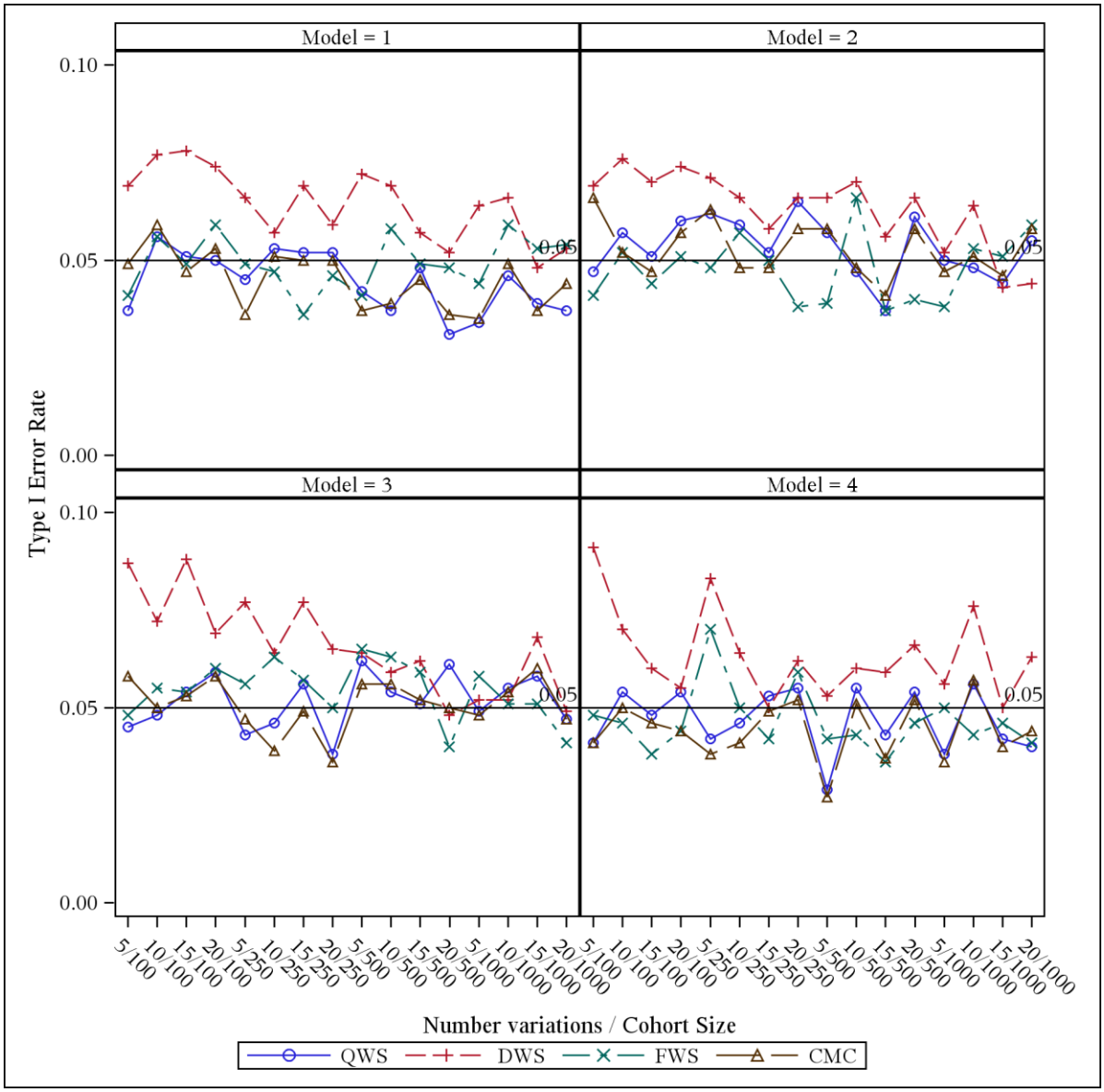

Figure 2.9: Type I error rates. Type I error rates for CMC, QWS, FWS and DWS on simulations with rare frequency variants.

References

Bodmer, W., & Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics, 40*(6), 695-701. doi:10.1038/ng.f.136

Cohen, J. C., Kiss, R. S., Pertsemlidis, A., Marcel, Y. L., McPherson, R., & Hobbs, H. H. (2004). Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science (New York, N.Y.), 305*(5685), 869-872. doi:10.1126/science.1099870

Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research, Saxena, R., Voight, B. F., Lyssenko, V., Burtt, N. P., de Bakker, P. I., et al. (2007). Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science (New York, N.Y.), 316*(5829), 1331-1336. doi:10.1126/science.1142358

Duerr, R. H., Taylor, K. D., Brant, S. R., Rioux, J. D., Silverberg, M. S., Daly, M. J., et al. (2006). A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science (New York, N.Y.), 314*(5804), 1461-1463. doi:10.1126/science.1135245

Fearnhead, N. S., Wilding, J. L., Winney, B., Tonks, S., Bartlett, S., Bicknell, D. C., et al. (2004). Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. *Proceedings of the National Academy of Sciences of the United States of America, 101*(45), 15992-15997. doi:10.1073/pnas.0407187101

Frazer, K. A., Murray, S. S., Schork, N. J., & Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nature Reviews.Genetics, 10*(4), 241-251. doi:10.1038/nrg2554

Gorlov, I. P., Gorlova, O. Y., Sunyaev, S. R., Spitz, M. R., & Amos, C. I. (2008). Shifting paradigm of association studies: Value of rare single-nucleotide polymorphisms. *American Journal of Human Genetics, 82*(1), 100-112. doi:10.1016/j.ajhg.2007.09.006

Gudmundsson, J., Sulem, P., Steinthorsdottir, V., Bergthorsson, J. T., Thorleifsson, G., Manolescu, A., et al. (2007). Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nature Genetics, 39*(8), 977-983. doi:10.1038/ng2062

Ji, W., Foo, J. N., O'Roak, B. J., Zhao, H., Larson, M. G., Simon, D. B., et al. (2008). Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nature Genetics, 40*(5), 592-599. doi:10.1038/ng.118

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., et al. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science (New York, N.Y.), 308*(5720), 385-389. doi:10.1126/science.1109557

Li, B., & Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *American Journal of Human Genetics, 83*(3), 311-321. doi:10.1016/j.ajhg.2008.06.024

Li, B., & Leal, S. M. (2009). Discovery of rare variants via sequencing: Implications for the design of complex trait association studies. *PLoS Genetics, 5*(5), e1000481. doi:10.1371/journal.pgen.1000481

Madsen, B. E., & Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics, 5*(2), e1000384. doi:10.1371/journal.pgen.1000384

Manolio, T. A., Brooks, L. D., & Collins, F. S. (2008). A HapMap harvest of insights into the genetics of common disease. *The Journal of Clinical Investigation, 118*(5), 1590-1605. doi:10.1172/JCI34772

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature, 461*(7265), 747-753. doi:10.1038/nature08494

Nejentsev, S., Walker, N., Riches, D., Egholm, M., & Todd, J. A. (2009). Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science (New York, N.Y.), 324*(5925), 387-389. doi:10.1126/science.1167728

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *American Journal of Human Genetics, 69*(1), 124-137. doi:10.1086/321272

Pritchard, J. K., & Cox, N. J. (2002). The allelic architecture of human disease genes: Common disease-common variant...or not? *Human Molecular Genetics, 11*(20), 2417-2423.

Reich, D. E., & Lander, E. S. (2001). On the allelic spectrum of human disease. *Trends in Genetics : TIG, 17*(9), 502-510.

Romeo, S., Pennacchio, L. A., Fu, Y., Boerwinkle, E., Tybjaerg-Hansen, A., Hobbs, H. H., et al. (2007). Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nature Genetics, 39*(4), 513-516. doi:10.1038/ng1984

Siva, N. (2008). 1000 genomes project. *Nature Biotechnology*, *26*(3), 256. doi:10.1038/nbt0308-256b

Todd, J. A., Walker, N. M., Cooper, J. D., Smyth, D. J., Downes, K., Plagnol, V., et al. (2007). Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature Genetics, 39*(7), 857-864. doi:10.1038/ng2068

Wray, N. R., & Visscher, P. M. (2010). Narrowing the boundaries of the genetic architecture of schizophrenia. *Schizophrenia Bulletin, 36*(1), 14-23. doi:10.1093/schbul/sbp137

Wright, S. (1949). Adaptation and selection. In *Genetics, Paleontology, and Evolution* (ed. G. L.Jepson, G.G.Simpson and E.Mayr). Princeton University Press.

CHAPTER 3

Application of Next Generation Sequencing to CEPH cell lines to discover variants associated with 30 FDA approved chemotherapeutics

Gunjan D. Hariani[1], Ernest Lamm[3], Tammy Havener[4], Pui Kwok[3], Howard L. McLeod[4], Michael J. Wagner[4], Alison A. Motsinger-Reif[1,2]

[1]Bioinformatics Research Center, North Carolina State University, Raleigh NC, USA;

[2]Department of Statistics, North Carolina State University, Raleigh NC, USA;

[3]Cardiovascular Research Institute, UCSF School of Medicine, San Francisco, CA, USA;

[4]Institute of Pharmacogenomics and Individualized Therapy; UNC Chapel Hill, NC, USA

Abstract

The goal of this study was candidate gene association with cytotoxicity of chemotherapeutics in cell line models through resequencing and discovery of rare and low frequency variants along with common variations. Here, an association study of cytotoxicity response to 30 FDA approved drugs was conducted and we applied next generation targeted sequencing technology to discover variants from 103 candidate genes in 95 lymphoblastoid cell lines from 14 CEPH pedigrees. In this article, we present the pipeline setup to robustly call variants across 95 cell lines and results from association analysis conducted with Family Based Association Testing software package (Laird, Horvath, & Xu, 2000). We called 1231 SNPs and 132 insertion/deletion markers across the 103 genes for these cell lines and identified three genes of significant association with this marker set. Specifically, ATP-binding cassette, sub-family C, member 5 (ABCC5) and NAD(P)H dehydrogenase, quinone1 NQO1 were significantly associated with oxaliplatin drug response. Interestingly, the significant SNP on NQO1 (rs1800566) has been linked with poor survival rates in patients with non-small cell lung cancer treated with cisplatin (which belongs to the same class of drugs as oxaliplatin) (Kolesar et al. 2011, 1765-1772). For the drug bleomycin, SNPs near 5' region of metallothionein (MT1A) were significant hits. The results from this study are promising and this serves as a proof-of-principle demonstration of the use of sequencing data in the cytotoxicity models of human cell lines. With increased sample sizes, such studies will

be a fast and powerful way to associate common and rare variants with drug response; while overcoming the cost and time limitations to recruit cohorts for association study.

Introduction

There is increasing evidence that genetic variation can explain some inter-individual variation in efficacy and toxicity across a spectrum of drugs used for treatment of various diseases. In situations where genetic factors dictate or at least predict drug response, individualized therapy holds a lot of promise and has already shown success in drug choice and dosing (Sim & Ingelman-Sundberg, 2011). A lot of these influential genetic factors have been identified by candidate gene studies but these studies require a priori knowledge about the involved genes. Despite the knowledge of essential genes involved in a drug metabolism pathway, these studies are generally difficult to conduct especially in pharmacogenomics due to difficulty in cohort recruitment, sample size limitations, phenotype characterization, time and cost factors, as well as replication of results in a different cohort (Daly, 2010; A. A. Motsinger-Reif et al., 2010).

An alternate approach to conducting candidate gene studies in clinical subjects is to use well established cell lines (e.g. lymphoblastoid cell lines, NCI60 cancer cell lines) for drug phenotyping and natural genetic variation in these cell lines for association testing in discovery phase. The use of cell line based model has its own limitations too – phenotypes may not completely reflect observed human phenotypes, not all enzymes involved in a drug response may be produced by the cell line and the choice of cell line for a study may bias the results. (For a review of cell line models in pharmacogenomics, refer Welsh et al., 2009). In spite of its limitations, this model has been used successfully to begin prioritization of 29 different US FDA approved chemotherapy drugs for further genetic analyses by estimating drug heritabilities using 125 lymphoblastoid cell lines (LCLs) from 14 CEPH pedigrees (Peters et al., 2011). This is a cost, time and resource effective strategy because exploring genetics only makes sense when it is contributing appreciably to a trait.

The robustness of this model for application to chemotherapy drugs has also been demonstrated by prioritizing candidate chromosomal regions within a class of drugs in an

agnostic approach. In a study by Watson and colleagues, the authors estimated heritability rates for six antitumor drugs (plus three replicates) belonging to the camptothecin class of drugs and conducted genome-wide linkage analysis with microsatellite markers (Watson et al., 2011). The six drugs had an average heritability estimate of 0.23 ± 0.026 and shared nine QTLs. Some QTLs were found in chromosomal regions known to influence the activity of camptothecins. The study also replicated the results in three other drugs belonging to the same class and nine out of ten shared QTLs were identified as significant in the replication set. In a similar study, Peters et al. established genome-wide linkage patterns for 29 chemotherapy drugs and reported overlap of QTLs within drug families Peters et al. (2011). A major challenge that arises from these studies is the selection of candidate genes from a myriad of genes that may harbor the significant QTL regions and find genetic variants for association.

As an alternative to the linkage-based approach described by Peters et al. (2011), we have undertaken an association study of cytotoxicity induced by chemotherapy drugs using targeted, next-generation sequencing to identify variants for testing. More than 100 candidate genes were targeted for sequencing in 95 CEPH cell lines from 14 pedigrees, and whole exome sequencing was also undertaken in 9 of the cell lines.

In this study, the power of next generation sequencing technologies has been harnessed to accurately identify novel and common variants in genes involved in pharmacodynamics and pharmacokinetics of cancer therapeutics (along with all exons in the genome). Family based association testing has been carried out via the FBAT software. Results from both the pilot whole exome study and candidate gene study are presented in this paper. After a multiple hypothesis correction was applied to each drug phenotype independently, we found statistically significant hits for two drugs (oxaliplatin and bleomycin) and some of the significant markers identified in this study have been previously implicated with drug response. This study demonstrates the potential of using resequencing in cell-line cytotoxicity models of drug response.

Materials and Methods

***Cytotoxicity Phenotypes*:** The cytotoxicity data from Peters et al were used as the phenotypes for association analysis. Briefly, 125 lymphoblastoid cell lines from 14 CEPH families were treated with four doses of each of 30 chemotherapy drugs (A list of drugs used in this study is available in Table 1) to capture the linear portion of the cell kill curve. Cell viability was used as a measure for drug induced cytotoxicity and was quantified using the non-toxic Almar Blue reagent which is converted into a fluorescent compound by the living cells. The fluorescence of drug treated cells was measured relative to cells treated with vehicle control (DMSO) to account for background noise. Outliers were removed and one average measurement across replicates was used for any further analysis. Details of the phenotyping experiments can be found in Peters et al (2011). Details of the quality control measures used can be found in A. A. Motsinger-Reif et al. (2011).Most of the drugs tested showed an estimated heritability of 0.3 or greater in the Peters et al. (2011) study indicating that the genetic component explains a significant amount of variability in drug response.

Selection of Cell lines for sequencing: Of the four drug concentrations for each antitumor agent, the one that resulted in an average viability closest to 50% across all cell lines was used for determining sensitivity or resistance to that agent. Cell lines demonstrating extreme responses, defined as viability below the 10^{th} or above the 90th percentiles of the viability distribution at the selected dose for a given drug, were respectively labeled as sensitive and resistant to that drug. Ninety-five CEPH cell lines that displayed sensitivity and/or resistance to at least one of 23 drugs were selected for candidate gene sequencing. The 23 drugs upon which the selection criteria were based included 14 drugs from one of 5 drug classes (fluoropyrimidines, anthracyclines, platinum compounds, taxanes, and camptothecins), as well as an additional 9 drugs for which the cytotoxicity profiles across the entire set of 125 cell lines showed high correlation with those of the 14 targeted drugs. A complete list of the 95 cell lines and pedigree structure is available in Appendix A/Table S.1. In addition, 9 cell lines were chosen for complete exome sequencing based on sensitivity to oxaliplatin: three

sensitive and six resistant cell lines comprising two trios and 3 unrelated individuals. Eight of these nine cell lines were also included in the group subjected to candidate gene sequencing.

***Candidate gene sequencing*:** 103 candidate genes were selected for resequencing based on their involvement in pathways for drug metabolism, transport, or drug action for 5 classes of chemotherapy drugs tested in the in vitro cytotoxicity assay described above: fluoropyrimidines, anthracyclines, platinum compounds, taxanes, and camptothecins. The candidate genes selected are listed in Appendix A/Table S.2. A multiplex PCR reagent for amplification of the exons, including untranslated regions and approximately 1000 bp upstream of the first exon, from all the candidate genes was designed by RainDance Technology (RDT). This technology allows for independent amplification of multiple PCR reactions in a single tube through the sequestration of primer pairs for each amplicon in separate, picoliter-volume microdroplets. (Tewhey et al., 2009) A total of 1932 amplicons were designed to capture the 103 candidate genes. The mean amplicon size was 514 bp (range 206-600 bp), and up to 18 amplicons were tiled to cover large exons. The total amount of genomic DNA sequence expected to be amplified by this PCR multiplex is 800,965 bp. After merging microdroplets containing genomic DNA from each of the 95 cell lines with the multiplex, primer microdroplet mix on an RDT1000 microfluidic station, samples were PCR amplified, amplicon DNA was purified, ligated, randomly sheared, and used to prepare sequencing libraries. Libraries were sequenced on Illumina Genome Analyzer II (GAIIx) sequencers to generate 36 base, single end sequences (i.e., only sequenced one end of the read) using between one and 9 sequencing lanes per sample, resulting in high depths and coverage across the amplicons.

***Whole exome sequencing*:** Whole exome capture was done on nine CEPH LCLs that included 2 trios and 3 unrelated cell lines. The exome capture was carried out using the SeqCap EZ Exome SR v1.1 hybridization capture reagent from Roche/NimbleGen, targeting approximately 28Mb of the genome. Library preparation and hybridization-based capture were carried out using the protocol specified in the SeqCap EZ Exome SR v1.1 user manual.

Paired end sequencing (i.e., sequence both ends of the read) was done for every sample on an Illumina GAIIx resulting in anywhere between 50-72 million reads of 75 bases in length per sample.

***Pipeline for variant discovery*:** The sequencing data obtained from different centers was subjected to rigorous data cleaning before variant calling. We received data for all samples in FASTQ format. A FASTQ is a standard file format to store sequence and Phred scaled base quality information (Cock, Fields, Goto, Heuer, & Rice, 2010). A Phred quality score is the -10*log10(estimated probability that the base was called incorrectly). Every FASTQ (141 for 95 samples) went through the following pipeline that was developed at Expression Analysis Inc, and modified for application to this dataset:

1. Adapter clipping and low quality base trimming: Adapters are known DNA sequences ligated to the fragmented DNA molecules that we want to sequence. These adapters allow the DNA molecule to attach to the flowcell (where the sequencing chemistry occurs) and also act as primers for cluster amplification (the DNA molecule is locally amplified on the flowcell to intensify the fluorescence signal when a base is incorporated during sequencing). Adapter clipping and read trimming were done using fastq-mcf (Aronesty, 2011).Any adapter sequences from the ends of the reads were clipped. Also, trailing low quality bases were trimmed from the ends of the reads. Both these steps improve the number of reads that can be mapped back to the genome. FASTQ quality statistics such as nucleotide distribution per cycle and base quality score per cycle were computed to make certain that the sequencing runs had not failed at any cycle and the FASTQs were of high quality.
2. Base Quality Score Recalibration: The Phred base qualities provided by Illumina are known to be inflated for higher quality values (DePristo et al., 2011) and can be corrected by incorporating the error rate from a PhiX control run. A heuristic polynomial model was used to correct for the Illumina provided base quality scores at the clipped FASTQ level. This was done to ensure only high quality variants were called in further analysis.

3. Read Alignment and pileup: Burrows-Wheeler Alignment Tool (BWA) v0.5.9 (Li & Durbin, 2009) was used to align reads to human genome (Hg19). Default parameters were used for alignment and generation of the Sequence Alignment Map (SAM) files that contain alignment coordinates and mapping qualities in a standard format (Li et al., 2009). Aligned reads were filtered for different thresholds of mapping qualities to test for effect on variant calling. Statistics for the aligned files were generated to gain an idea of the capture quality by quantifying the number of reads mapped, mapping qualities, and percent alignment on different chromosomes. The SAM files were then converted to pileup formats for quality statistics using Samtools v0.1.15 (Li et al., 2009). Pileup format gives base-by-base information for all aligned reads in terms of chromosomal position, reference base, number of reads aligned at that position, base calls, and base qualities (Li et al., 2009). Rigorous quality statistics were computed using the pileup format to assess the PCR and hybrid capture processes. These quality checks include coverage, depth, uniformity of coverage, number of bases that fall inside of amplicons vs. outside of amplicons, number of reads that fall within the amplicons vs. outside of them (unpublished work, Expression Analysis Inc, Durham), and total number of amplicons captured.
4. Variant Calling: Prior to SNP calling, samples with replicate runs were merged into one BAM file (binary equivalent of SAM). Along with the merged bam file, variant calling was conducted on each replicate to allow for consistency check. The bam files were then realigned around regions containing insertions or deletions (indels) to minimize the number of mismatch SNPs and false SNP calls and used for single sample SNP calling. For indel calling, a secondary realignment around known dbSNP indel regions (Available from: http://www.ncbi.nlm.nih.gov/SNP/) was done, followed by variant calling. The realignment and variant calling were done using the open source JAVA software, Genome Analysis Tool Kit (version 1.4-30-gf2ef8d1) (GATK) (McKenna et al., 2010) using the default parameters and all available depth (no downsampling was carried out). Genotype calls provided by GATK were used for association analysis.

5. Variant Filtering: Due to the few number of variant calls made in the candidate regions, the GATK variant filtering could not be used; and custom filters were set up to ensure that highest quality variants passed through. SNPs meeting any of the following criteria were masked to unknown genotypes:
    a. Quality By Depth (GATK parameter): <=2.5; This is a GATK parameter which is computed as phred scaled probability of observing a variant at the given site/depth at the site Low QD values may indicate errors (Broad Institute, 2011)
    b. Strand Bias (GATK parameter): >=60; GATK parameter where high values indicate that mostly one of the strands is showing evidence for the variant.
    c. Depth: <=5
    d. Known dbSNP indel was called within the individual at this position
    e. If any two of the following conditions were met, the genotype calls were masked out:
        i. The site under consideration was called with multiple alleles in the given sample of 95 individuals
        ii. Additional variant calls were made within a few bases of the site in this individual or the population (Indicator of alignment error or sequencing error of low complexity region)
        iii. The variant call was made within 3 bases to the left of the sense primer or 3 bases to the right of antisense primer
        iv. Mendelian rules of transmission were used to rescue inaccurately called genotypes in offspring – this affected no more than one or two calls within an individual

Recommendations from (DePristo et al., 2011) were used for indel filtering. Indels that had any of the following quality statistics were masked out:

a. Quality By Depth: <2
b. Homopolymer Run: >=7
c. %Reads with mapping quality zero: >5

6. Variant Quality/Genotype Quality assessment: Genotypes from samples were compared to available Hapmap phase III sample calls (http://hapmap.ncbi.nlm.nih.gov/). Genotypes across different replicates in candidate gene study were compared; and genotypes between candidate gene study and exome study data were contrasted. Genotype consistency was also determined by Mendelian error checks. To compute Mendelian error, a superset of variant positions was generated for every nuclear family (trio) in a pedigree. For each variant position, if a variant call was not made in a family member in presence of sufficient depth (20), then a homozygous reference genotype was assigned to that member. Variant Quality assessment was also made in terms of dbSNP membership.

***Association analysis*:** Family Based Association Testing (FBAT version 2.0.3) software (Laird et al., 2000) was conducted to test the null hypothesis of no linkage or no association of marker with unknown trait locus. The testing was conducted for autosomal markers only (sex linked markers were not used in this study). The minimum number of informative families was set to 4 in order to maximize the utilization of available marker set. The offset value for each phenotype was set to the sample mean of that phenotype because the phenotypes were relatively normal and no ascertainment bias from extreme percentiles of cell viability distribution was observed. We also conducted association of Glutathione S-transferase Mu 1 (GTM1) gene deletion with cell line viabilities. GSTM1 belongs to a class of enzymes called Glutathione S-transferase that are involved in detoxification of therapeutic drugs and has been implicated in chemotheraphy efficacy (Mossallam, Abdel Hamid, & Samra, 2006), (Voso et al., 2002). GSTM1 gene deletion is known to occur in >75% of Hapmap CEU samples (Huang et al., 2009).We inferred homozygous gene deletion for GSTM1 targeted regions using coverage statistics and fit a mixed effects model for each drug (families where all members showed gene deletion or gene presence were not used which left us with 7 families). The following model was fit:

$$\text{Phenotype}(ijk) = \mu + \alpha(i) + F(j) + \varepsilon(ijk) \qquad (1)$$

μ = reference level, $\alpha(i)$ = fixed effect of GSTM1 homozygous deletion ($i = 0,1$), $F(j)$ = random effect due to family ($j=1..7$), phenotype(ijk) = phenotype observation for individual k belonging to family j with deletion status i (0 = no homozygous deletion, 1 = homozygous

deletion). We were interested in testing the null hypothesis that there is no effect of homozygous deletion, i.e., H_o: $\alpha = 0$; H_a: $\alpha \neq 0$.

Multiple Hypothesis Correction: We applied FDR correction (Benjamini & Hochberg, 1995) for the number of markers informative in at least four families – correction was applied independently to every drug to account for multiple comparisons and control the family-wise error rate at α=0.05.

Pipeline Implementation: The pipeline for variant calling was implemented in bash scripting. Variant filtering was done using custom R and perl scripts. The pipeline was run on a cluster of 30 commodity servers with 4 to 24 CPUs per node, 8 to 60 GB of RAM per node, and Ubuntu Linux as the operating system. To process a single FASTQ with approximately 14M reads for steps 1 – 3 on a single CPU with 40 GB of RAM required 2 hours and 45 min of CPU time. Alignment to the human genome was the most time consuming step (91 minutes).

Results

Candidate Gene Study:

***FASTQ quality assessment*:** Due to processing of DNA samples at four sequencing centers, we had 141 FASTQ files (FASTQs) generated for 95 LCLs with each sample having between 1 to 9 FASTQs (for the candidate gene study). All FASTQs were assessed for quality control by examining base distribution per cycle and quality by cycle. The nucleotide base distribution was even for every cycle indicating no cycle bias and no primer bias as is expected for DNA sequencing. The mean base qualities across FASTQs began to drop with increments in sequencing cycle, but the drop is very mild (Appendix A/Figure S.1). The read lengths and number of reads per FASTQ varied for different samples across centers. The number of reads per FASTQ ranged from 16K to 29M indicating cluster generation problems for certain flowcells (Appendix A/Table S.3, Appendix A/Figure S.2). The low count read

FASTQs from RDT run 5 were discarded from any further analysis as this entire flowcell showed cluster generation problems. We were left with 134 FASTQs for the 95 samples.

***SAM quality and Amplicon capture assessment*:** The FASTQs were processed and aligned to UCSC hg19 assembly using BWA software. For the candidate gene study, 94.08% reads on average (± 1.52% sd) aligned per sample with a median mapping quality score of 37. The percent distribution of mapped bases per chromosome was comparable to the expected distribution; although the variability in capture indicates the possibility of high genomic contamination, variability in FASTQ read counts from different sequencing centers, and failure of few primers across samples at library preparation level (Appendix A/Figure S.3A).

Samples with higher number of reads showed higher coverage – more targeted amplicons were sequenced and at higher mean depths. The average depth over all amplicons in a sample was 292.23 (± 148.74 sd). The coverage (percent bases across all amplicons covered at least 1X) varied from 87.33 – 99.85%. The high variability in mean depth and coverage can be accounted to low read counts (RDT Runs 3, 4 and 6) and PCR capture differences per sample (Solexa Run 6) (Appendix A/Figure S.4). We also assayed specificity of primer performance in terms of percent bases on target vs. outside of targeted regions. The percent of bases on target was low given that the capture was meant to be very specific, and it ranged from 15 to 40% across samples (Appendix A/Figure S.5). We tested to see if the primers align ambiguously to the human genome resulting in low capture specificity. Out of approximate 1930 primer pairs, BWA aligned 493 primers (12.7%) with zero mapping quality. The ambiguous alignment of a small proportion of primers cannot explain the high percentage of bases outside of targeted regions. Given that most primers were unique and the targeted regions had excellent coverage, we do not believe this was a primer-specific issue and present support for possibility of high genomic contamination. The quality of bases within regions of interest (ROI) were checked in terms of read mapping quality – on median across 134 samples, 5.06% bases in ROI came from reads with mapping quality zero. This ensures that most of the targeted region was suitable for faithful variant calling.

The uniformity of capture at a given depth was calculated for every amplicon per sample using a measure called Area Under the Reference Line (AURL) (unpublished work, Expression Analysis Inc, Durham), where the reference line is defined by the depth in consideration. For a uniform coverage at a certain depth, the AURL will be close to 1; non-uniformity and depths lower than reference line reduce the value of AURL towards zero. The average AURLs at depths 20, 30, 50 and 100 are plotted in Appendix A/Figure S.7. Overall, the mean AURL within a sample was 95.90 and 94.41% at depths 20 and 30 respectively; signifying that the amplicons within an individual were uniformly covered at sufficient depth to call variants for most positions. The samples that showed low AURLs belonged to RDT Runs 3, 4 and 6. Given the relatively low performance of lanes from RDT Runs 3, 4 and 6; 15 replicate samples were discarded from further analysis without resulting in any data loss for the 95 LCLs.

Every gene was assessed for primer capture efficiency (Appendix A/Figure S.6). Two genes GSTM1 on chromosome 1 and TYMS on chromosome 18 showed low capture proficiency. All amplicons for GSTM1 failed to be captured in some samples; whereas only 1 amplicon for TYMS was non-uniformly captured across few samples. This shows that primer specific capture of RDT worked for most amplicons across all samples.

Exome Data Study:

***FASTQ quality assessment*:** For the whole exome data, all the sequencing for 9 samples was done at one center. No anomalies were found in any of the FASTQs indicating that the sequencing runs had completed successfully without errors. The number of read pairs varied from 25M to 36M per sample (Appendix A/Table S.3). Lower read duplicates were detected for exome sequencing (0.04%-2.36%) than candidate gene study (on average 17% reads were duplicated) due to increased complexity of capture regions for the exome study.

***SAM quality Assessment*:** We observed a higher mapping percentage of reads for exome study as compared to the candidate gene study (98.49 ± 0.72). The mapping qualities (median = 60) for this study are also higher than the candidate gene study. Both of these can

be contributed to longer read lengths (75) and availability of paired end reads. There was little variability in percentage of reads mapping to chromosomes across different samples as compared to the RDT runs.

Variant Calling:

Genome Analysis Toolkit from BROAD Institute was used for variant calling and filters described in Methods section were set to retain high quality variants. Appendix A/Figure S.8 summarizes Mendelian segregation error rates for the trios generated from all pedigrees after SNP variant filtration. One of the family trio (offspring 12142 from pedigree 1334) showed very high Mendelian error rate (39.19%) – indicating that the sample might have been mislabeled and was excluded from any further analysis. The remaining trios showed on an average 10.09 ± 6.047% error rates before filtering and 0.995 ± 0.696% after filtering. Hapmap data was available for 18 samples used in this study – this was used to check genotype consistency from sequencing data. Before filtering, an average of 96.12 ± 2.86% of variants in Hapmap were called in sequencing data with matching genotypes and 99.82 ± 0.23% of homozygous reference calls were called non-variant sites in sequencing data. After filtering, an average of 95.06 ± 3.38% of variants in Hapmap were called consistently in sequencing data and 100% of reference calls were called non-variant in sequencing data (we improved specificity as the cost of sensitivity) (Appendix A/Figure S.8). Variant quality was further assessed by looking at dbSNP membership of the calls. Most of the filtered calls were non-dbSNP members (Appendix A/Figure S.8). The various quality assessments ensure that we were left with 1231 high quality SNPs in the targeted regions across all cell lines for association analysis. Genotype calls for >78% SNPs were available in 90 samples or more (Appendix A/Figure S.9).

GATK pipeline was used to do local realignment around known dbSNP indels and call indels in individual samples. The default downsampling parameter (by default the coverage is cut off to 250) was over-ridden to no downsampling as it resulted in biased indel genotype calling – this maybe due to uneven sampling of reads. We discovered 132 known dbSNP indels across the 94 cell lines. Genotypes at indel positions were available for >55%

of markers in 90 samples or more. We observed high Mendelian errors for indel calling (a total of 121 Mendelian errors were found across all trios for 132 indel positions as compared to 136 Mendelian errors observed across all trios for 1231 SNP positions). This is not surprising as differentiating alignment errors/sequencing errors as a result of homopolymer runs from indels is a difficult problem.

Exome and Candidate study comparison:

Due to the limited number of samples (9) chosen for the exome data study; we could not use this dataset for association analysis. Instead, we used it to examine genotype calling fidelity from the candidate regions sequencing. The exome sequencing study had a low uniform coverage at depth 20 for the regions studied in candidate gene study. Hence, a lot of calls made in the candidate study were not observed in the exome dataset. For the variant positions called in common between the two studies, we observed a 100% match for variant calling across all 9 individuals (i.e. the same variant allele was called at all positions in both the studies). Discrepancy was observed in terms of genotype assignment; most of which may be attributed to differences in sequencing depth. One of the nine samples (12765) showed a very high rate of genotype discrepancy (33%) as compared to other samples (0 – 0.9%). It is unclear if this is due to sequencing of depth bias, differences in capture protocol for the two studies, or any other systematic bias.

FDR Results:

The 94 cell lines used in this association study were chosen such that each cell line was an extreme responder (sensitive or resistant) to at least 1 drug. Despite this design of selecting responders from extreme percentiles, each drug had responders spanning across different percentiles of viability. There was no ascertainment bias for any given drug and we chose an offset of sample phenotypic mean to maximize the power of FBAT as recommended in (Lange, DeMeo, & Laird, 2002). We applied FDR correction for the number of markers informative in at least four families – correction was applied independently to every drug. Two drugs (Bleomycin and Oxaliplatin) had statistically

significant hits at a FDR corrected threshold of 0.05. Two known SNPs upstream of gene metallothionein (MT1A) on chromosome 16 are associated with bleomycin (Figure 1), whereas five common known SNPs - 2 synonymous and 3 in UTR3 region of ATP-binding cassette, sub-family C, member 5 (ABCC5/MRP5) gene on chromosome 3 were significantly associated with oxaliplatin (Figure 2). Additionally, a SNP in NAD(P)H dehydrogenase, quinone1 (NQO1) gene on chromosome 16 (rs1800566) of clinical significance in resistance to other chemotherapy drugs (http://www.pharmgkb.org/) is also in association with oxaliplatin. Details can be found in Table 2 and Appendix A/Table S.4. Linkage disequilibrium (LD) (r^2) was calculated and haplotype inference was conducted to assess correlation between SNPs using Haploview ((Barrett, Fry, Maller, & Daly, 2005)). SNP markers discovered in the 94 samples were used for LD and haplotype inference. Default parameters were used. From the markers used, we observe that the ABCC5 gene on chromosome 3 can be broken down into 2 haplotype blocks interspersed with low frequency alleles in our samples (MAF < 0.05; as estimated in Haploview from the subset of samples used to generate the LD blocks) (Appendix A/Figure S.10). Functional annotations on the 8 low frequency alleles reveal that 2 are exonic (benign) (Adzhubei et al., 2010); 4 are in UTR3 and the remaining 2 are in UTR5 regions of ABCC5 (K. Wang, Li, & Hakonarson, 2010). For the significant hits on chromosome 16 with bleomycin, we observe that the two significant hits are in perfect correlation with each other (r^2=1) but not contained within any haplotype block (Appendix A/Figure S.11). Five low frequent alleles are observed within 2Kb regions of the significant hits – no functional annotation was performed as this entire region is contained outside of genic regions.

Association results with GSTM1 deletion:

We used sequencing information to determine deletion of GSTM1 in individuals from amplicon capture assessment. Sixty-two samples showed no coverage across at least 4 of 6 targeted regions in GSTM1 at depth 20 and were determined to have homozygous deletion for GSTM1 (Appendix A/Figure S.6B). A Transmission Disequilibrium (TDT) form of test (where over-transmission of allele from heterozygous parents to affected offspring is

accessed (Spielman, McGinnis, & Ewens, 1993) was not done because it was not possible to determine if the carrier of GSTM1 had one or two copies of the gene and hence genotype assignment was not available. Instead, we fit a mixed effects model as explained in methods section to assay differences in drug response due to GSTM1 deletion. No drug showed statistically significant differences in cell viability across individuals due to homozygous deletion of GSTM1 gene after accounting for family-to-family variability (FDR cutoff = 0.05 was used).

Discussion and Conclusions

This study selected 14 CEPH LCL pedigrees to perform association analysis of 103 candidate genes to drug response for 30 chemotherapeutics. The 103 candidate regions were sequenced with sufficient depth to determine variants within each LCL. Sequencing data allowed us to have a higher resolution for the markers called than genotyping data – we had more variant calls of high quality than Hapmap dataset for the 18 samples in common. We detected variants present at low frequency (<5%) in the sample data, made indel calls and inferred homozygous gene deletion in samples. The added benefit of detecting more variants than offered by a genotyping chip comes at the cost of time required to set up the pipeline for variant calling (even with open source software), computational time, space and resources required to process each sample and data management. Approximately 2.5TB of data generated from this sequencing study (including raw FASTQs, processed FASTQs, intermediate alignment files, and variant call files) allowed us to call 1231 SNP markers and 132 indel markers across 94 samples.

We conducted FBAT for all the markers discovered in this study for 30 drug response phenotypes. When accounting for all markers and drugs, no hits are significant after an FDR correction. This is not surprising and the loss of power to detect any significant association can be attributed to the small sample size used in this study and a heavy burden of multiple hypotheses testing correction (for 1363 markers and 30 drug phenotypes. However, when we do not apply correction for phenotypes but only correct for multiple markers per drug, we observe significant hits for 2 drugs at FDR corrected threshold of 0.05. Five common SNPs

on ABCC5 gene (three in 3` UTR region and two in exons) are significantly associated with oxaliplatin (platinum-based drug) response. ABCC5 belongs to the class of ATP-binding cassette transporters and is involved in cellular transport of cyclic nucleotides (RefSeq, 2008). ABCC5 has previously been associated with oxaliplatin resistance (Pratt et al., 2005) and other platinum drugs ((Oguri et al., 2000)); although the role of ABBC5 in drug resistance is not clear (Borst, de Wolf, & van de Wetering, 2007). In addition to ABCC5, a non-synonymous variant (rs1800566) in NQO1 was associated with oxaliplatin. NQO1 is a member of the platinum pathway and is involved in metabolism of platinum based drugs. NQO1 has been linked with resistance to another platinum containing drug cisplatin (Kolesar et al., 2011), (Cho, Manandhar, Lee, Park, & Kwak, 2008), (X. J. Wang et al., 2008) and may share a common role with metabolism of oxaliplatin. The other drug that showed significant hits is bleomycin and showed statistical significance with 2 snps near 5' of MT1A gene. Haploview analysis in ±2Kb region around these SNPs containing genes MT1A and MT1JP did not reveal any correlation with other SNPs. Uncharacterized variants have been associated with bleomycin in this study and maybe valuable in unraveling the less understood mechanism of bleomycin.

While only 2 of 30 drugs showed significant hits, it is quite possible that none of the markers had a big enough effect size for other drugs to be easily detected in this study. Also, it is well known that drug response is a complex phenotype and SNP interactions may be able to explain some of the observed variability in drug response; but they were not tested in this study. There is a possibility that mechanisms other than polymorphisms in genes may have an influence on drug phenotype (e.g. miRNA, transcript expression).In conclusion, while the small sample size limits the inference that can be made in the current data, this study is a proof-of-principle demonstration of the use of sequencing data in the cytotoxicity models of human cell lines. As sequencing data becomes more accessible, such an approach will likely be more commonly applied to associate rare and novel variants alongside common variants with drug response that would have otherwise been missed by GWAS chips.

Acknowledgments

This work is supported by the NIH U01 GM63340 and NCI R01 CA161608 grants. We would also like to thank Victor J. Weigman and Wendell D. Jones from Expression Analysis Inc, Durham, NC 27713 for help with pipeline setup, useful discussions and comments on this paper.

Table3. 1: List of 30 drugs used in this study. The drugs and their chemical structure/functional class are listed.

Drug	Class
Busulfan	alkyl sulfonate
Mitoxantrone	anthracenediones
Daunorubicin	anthracyclines
Doxorubicin	anthracyclines
Epirubicin	anthracyclines
Idarubicin	anthracyclines
Topotecan	camptotheca
Hydroxyurea	deoxyribonucleotide
Trichostatin A	histone deacetylase inhibitor
Rapamycin	mTOR inhibitor
Carboplatin	platinum
Oxaliplatin	platinum
Etoposide	podophyllum
Teniposide	podophyllum
Cladribine	purine
Fludarabine	purine
5-fluorouracil	pyrimidine
Azacitidine	pyrimidine
Cytarabine	pyrimidine
Floxuridine	pyrimidine
Gemcitabine	pyrimidine
Bleomycin	streptomyces
Mitomycin	streptomyces
Docetaxel	taxane
Paclitaxel	taxane
Temozolomide	triazines
Vinblastine	vinca alkaloid
Vincristine	vinca alkaloid
Vinorelbine	vinca alkaloid
Arsenic trioxide	Other

Table 3.2: Significant hits from association. Significant hits from association at FDR corrected threshold of 0.05 are shown in this table. All hits are known dbSNP common variants. The allele frequency is estimated in the low coverage pilot study of 1000 Genomes Project in CEU samples (n=120). All data is reported from Database of Single Nucleotide Polymorphisms (build 136).

Drug(s)	**SNP ID**	**SNP Position**	**Gene**	**CEU Frequency**	**q-value**
Bleomycin	rs60900828 (A/T)	chr16:56671632	MT1A (near-gene 5)	A:0.875/ T:0.125	0.048755
	rs59104702 (A/C)	chr16:56671717	MT1A (near-gene 5)	A:0.108/ T:0.892	0.048755
Oxaliplatin	rs1800566 (C/T => P187S)	chr16:69745145	NQO1 (exonic)	C:0.783/ T:0.217	0.04116
	rs562 (C/T)	chr3:183637845	ABCC5 (UTR3)	C:0.517/ T:0.483	0.032536
	rs3749445 (A/G)	chr3:183638506	ABCC5 (UTR3)	A:0.508/ G:0.492	0.032536
	rs939336 (A/G => C594C)	chr3:183685534	ABCC5 (exonic)	A:0.408/ G:0.592	0.032536
	rs1132776 (C/T => A395A)	chr3:183696402	ABCC5 (exonic)	C:0.575/ T:0.425	0.032536
	rs4148575 (C/T)	chr3:183702275	ABCC5 (UTR3)	C:0.592/ T:0.408	0.032536

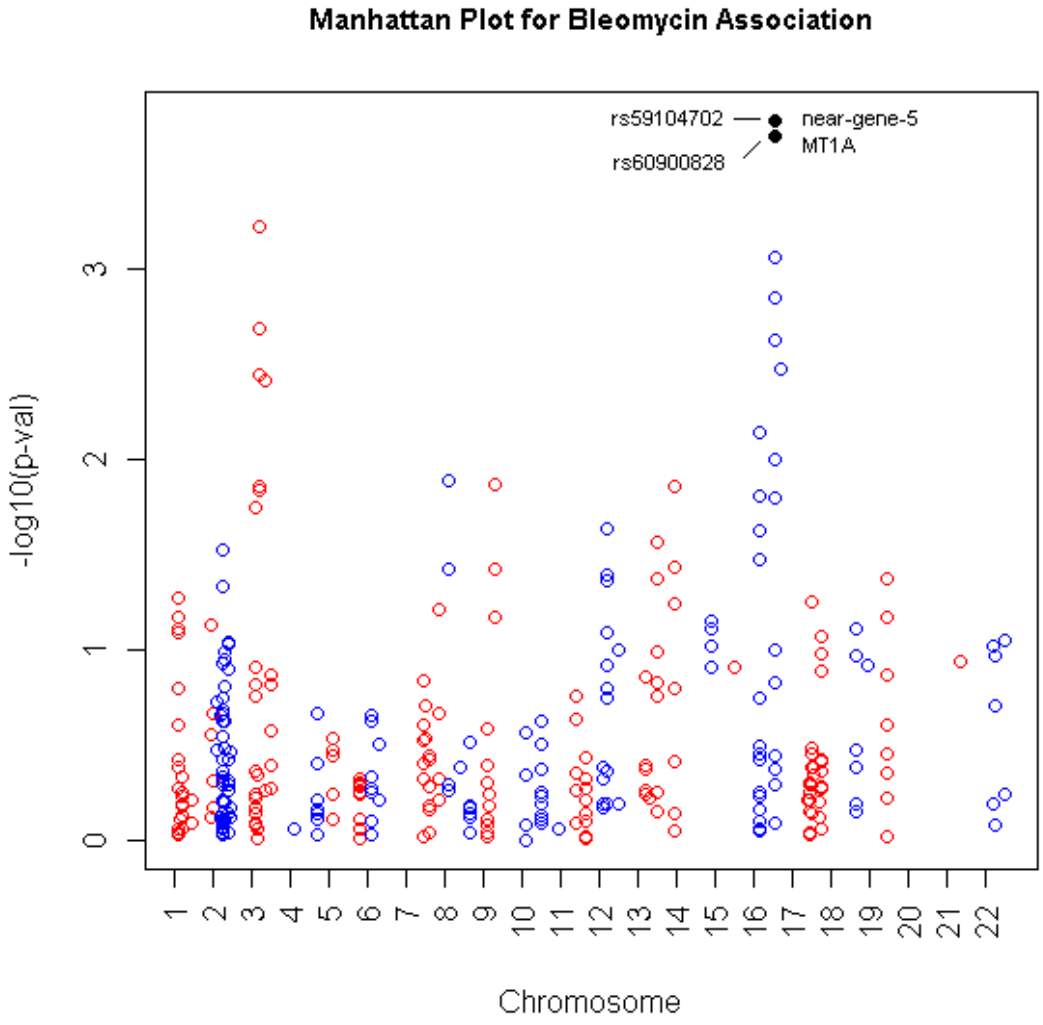

Figure 3.1: Manhattan plot for association with Bleomycin. Manhattan plot for association with Bleomycin is shown. The markers in black are significantly associated at a FDR corrected threshold of 0.05.

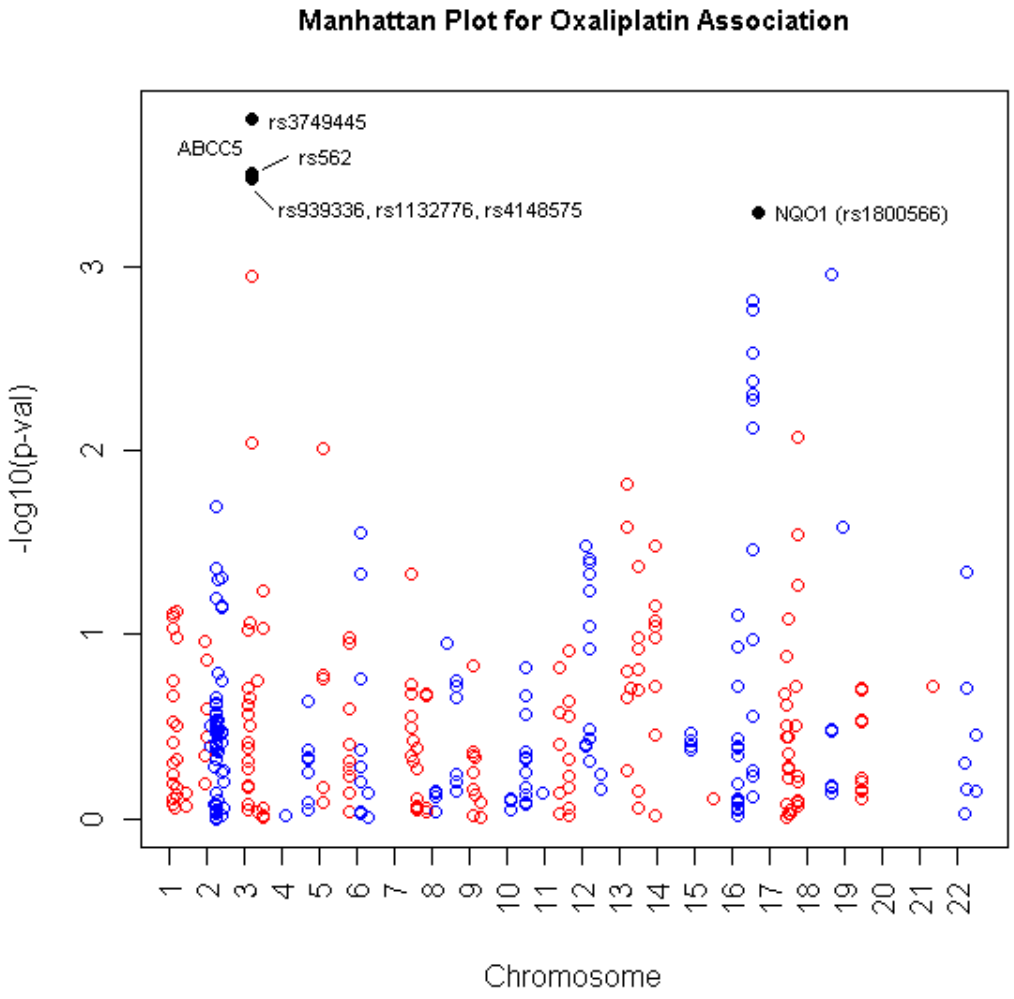

Figure 3.2: Manhattan plot for association with Oxaliplatin. Manhattan plot for oxaliplatin is shown. The markers in black are significantly associated at a FDR corrected threshold of 0.05.

References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. *Nature Methods, 7*(4), 248-249. doi:10.1038/nmeth0410-248

Aronesty, E. (2011). *Command-line tools for processing biological sequencing data*

Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2005). Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics (Oxford, England), 21*(2), 263-265. doi:10.1093/bioinformatics/bth457

Borst, P., de Wolf, C., & van de Wetering, K. (2007). Multidrug resistance-associated proteins 3, 4, and 5. *Pflügers Archiv European Journal of Physiology, DOI 10.1007/s00424-006-0054-9*(453), 661.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series* B, 57, 289-300.

Broad Institute. (2011). *Understanding the unified genotyper's VCF files.* Retrieved June 16, 2012, from http://www.broadinstitute.org/gsa/wiki/index.php/Understanding_the_Unified_Genotyper%27s_VCF_files

Cho, J. M., Manandhar, S., Lee, H. R., Park, H. M., & Kwak, M. K. (2008). Role of the Nrf2-antioxidant system in cytotoxicity mediated by anticancer cisplatin: Implication to cancer cell resistance. *Cancer Letters, 260*(1-2), 96-108. doi:10.1016/j.canlet.2007.10.022

Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research, 38*(6), 1767-1771. doi:10.1093/nar/gkp1137

Daly, A. K. (2010). Genome-wide association studies in pharmacogenomics. *Nature Reviews.Genetics, 11*(4), 241-246. doi:10.1038/nrg2751

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics, 43*(5), 491-498. doi:10.1038/ng.806

Huang, R. S., Chen, P., Wisel, S., Duan, S., Zhang, W., Cook, E. H., et al. (2009). Population-specific GSTM1 copy number variation. *Human Molecular Genetics, 18*(2), 366-372. doi:10.1093/hmg/ddn345

Kolesar, J. M., Dahlberg, S. E., Marsh, S., McLeod, H. L., Johnson, D. H., Keller, S. M., et al. (2011). The NQO1*2/*2 polymorphism is associated with poor overall survival in patients following resection of stages II and IIIa non-small cell lung cancer. *Oncology Reports, 25*(6), 1765-1772. doi:10.3892/or.2011.1249; 10.3892/or.2011.1249

Laird, N. M., Horvath, S., & Xu, X. (2000). Implementing a unified approach to family-based tests of association. *Genetic Epidemiology, 19 Suppl 1*, S36-42. doi:2-M

Lange, C., DeMeo, D. L., & Laird, N. M. (2002). Power and design considerations for a general class of family-based association tests: Quantitative traits. *American Journal of Human Genetics, 71*(6), 1330-1341. doi:10.1086/344696

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics (Oxford, England), 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England), 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research, 20*(9), 1297-1303. doi:10.1101/gr.107524.110

Mossallam, G. I., Abdel Hamid, T. M., & Samra, M. A. (2006). Glutathione S-transferase GSTM1 and GSTT1 polymorphisms in adult acute myeloid leukemia; its impact on toxicity and response to chemotherapy. *Journal of the Egyptian National Cancer Institute, 18*(3), 264-273.

Motsinger-Reif, A. A., Brown, C., Havener, T., Hardison, N. E., Peters, E. J., Beam, A., et al. (2011). Ex-vivo modeling for heritability assessment and genetic mapping in pharmacogenomics.

Motsinger-Reif, A. A., Jorgenson, E., Relling, M. V., Kroetz, D. L., Weinshilboum, R., Cox, N. J., et al. (2010). Genome-wide association studies in pharmacogenomics: Successes and lessons. *Pharmacogenetics and Genomics*, doi:10.1097/FPC.0b013e32833d7b45

Oguri, T., Isobe, T., Suzuki, T., Nishio, K., Fujiwara, Y., Katoh, O., et al. (2000). Increased expression of the MRP5 gene is associated with exposure to platinum drugs in lung cancer. *International Journal of Cancer.Journal International Du Cancer, 86*(1), 95-100.

Peters, E. J., Motsinger-Reif, A., Havener, T. M., Everitt, L., Hardison, N. E., Watson, V. G., et al. (2011). Pharmacogenomic characterization of US FDA-approved cytotoxic drugs. *Pharmacogenomics, 12*(10), 1407-1415. doi:10.2217/pgs.11.92

Pratt, S., Shepard, R. L., Kandasamy, R. A., Johnston, P. A., Perry, W.,3rd, & Dantzig, A. H. (2005). The multidrug resistance protein 5 (ABCC5) confers resistance to 5-fluorouracil and transports its monophosphorylated metabolites. *Molecular Cancer Therapeutics, 4*(5), 855-863. doi:10.1158/1535-7163.MCT-04-0291

RefSeq. (2008). *RefSeq gene ABCC5.* Retrieved June 22, 2012, from https://cgwb.nci.nih.gov/cgi-bin/hgc?hgsid=111190&g=refGene&i=NM_005688&c=chr3&o=183637725&l=183637725&r=183735727&db=hg19

Sim, S. C., & Ingelman-Sundberg, M. (2011). Pharmacogenomic biomarkers: New tools in current and future drug therapy. *Trends in Pharmacological Sciences, 32*(2), 72-81. doi:10.1016/j.tips.2010.11.008

Spielman, R. S., McGinnis, R. E., & Ewens, W. J. (1993). Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics, 52*(3), 506-516.

Tewhey, R., Warner, J. B., Nakano, M., Libby, B., Medkova, M., David, P. H., et al. (2009). Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nature Biotechnology, 27*(11), 1025-1031. doi:10.1038/nbt.1583

Voso, M. T., D'Alo', F., Putzulu, R., Mele, L., Scardocci, A., Chiusolo, P., et al. (2002). Negative prognostic value of glutathione S-transferase (GSTM1 and GSTT1) deletions in adult acute myeloid leukemia. *Blood, 100*(8), 2703-2707. doi:10.1182/blood.V100.8.2703

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research, 38*(16), e164. doi:10.1093/nar/gkq603

Wang, X. J., Sun, Z., Villeneuve, N. F., Zhang, S., Zhao, F., Li, Y., et al. (2008). Nrf2 enhances resistance of cancer cells to chemotherapeutic drugs, the dark side of Nrf2. *Carcinogenesis, 29*(6), 1235-1243. doi:10.1093/carcin/bgn095

Watson, V. G., Motsinger-Reif, A., Hardison, N. E., Peters, E. J., Havener, T. M., Everitt, L., et al. (2011). Identification and replication of loci involved in camptothecin-induced cytotoxicity using CEPH pedigrees. *PloS One, 6*(5), e17561. doi:10.1371/journal.pone.0017561

Welsh, M., Mangravite, L., Medina, M. W., Tantisira, K., Zhang, W., Huang, R. S., et al. (2009). Pharmacogenomic discovery using cell-based models. *Pharmacological Reviews, 61*(4), 413-429. doi:10.1124/pr.109.001461

CHAPTER 4

Application of sequencing methods to refine an association signal identified in a preliminary study for small for gestation age births

Gunjan D. Hariani[1], Shengdar Tsai[3,4], Alison A. Motsinger-Reif[1,2], Jorge A. Piedrahita[3,4]

[1]Bioinformatics Research Center, North Carolina State University, Raleigh NC, USA; [2]Department of Statistics, North Carolina State University, Raleigh NC, USA; [3]Department of Molecular Biomedical Sciences, College of Veterinary Medicine, North Carolina State University, Raleigh, NC, USA; [4]Center for Comparative Medicine and Translational Research, North Carolina State University, Raleigh, NC, USA

Abstract

Small for gestation age (SGA) is a pregnancy related disorder where the baby is smaller than 90% of the babies at the same gestational age. A genome wide transcription analysis study conducted in 63 human placentas (Caucasians and African Americans) found that the expression of an imprinted gene (paternally expressed 3 – PEG3) was significantly different in SGA and controls. A follow-up study with two common markers in PEG3 established an association of a common SNP marker rs1055359 in the UTR3 region of PEG3 with SGA in 158 Caucasian patients but not in African American samples (n=123).
The goal of this study was to investigate the allelic architecture (any evidence of rare variants) of PEG3 as well as the functional role of SNPs in the region co-located around rs1055359. Fifty-five placental samples (23 Caucasian and 32 African American) were chosen for resequencing a 0.5Mb region around PEG3. The SNPs found in PEG3 for these samples were functionally annotated and evaluated for allele frequency. Association analysis was conducted to test for accumulation of low frequency variants in SGA vs. controls. No significant results for accruement of low frequency variants in exonic region of PEG3 were observed for SGA vs. control samples in both ethnicities suggesting that the original signal maybe of most significance.

Introduction

This study pertains to a pregnancy related trait called small for gestational age (SGA); where an infant that weighs less than 90% of the babies at the same gestational age is classified as SGA (after accounting for other factors like ethnicity and gender). SGA infants have a higher risk for mortality, restricted growth and development (CDC, 2008). As per a study conducted on singleton birth data across USA from 1995-1999, it was estimated that 10.5% of the live births were SGA. The estimation for infant death rate was 5.2 and 16.7 per 1000 live births for non-SGA and SGA respectively (Kristensen et al., 2007).

The authors of this study, Tsai et al., were interested in the placental factors that could potentially account for restricted growth. Through a preliminary transcriptional profiling study done in 63 samples, the authors found that the expression level of the paternally expressed 3 (PEG3) gene was significantly different in the SGA and non-SGA groups (at FDR corrected threshold of 0.10). This finding was interesting as PEG3 is an imprinted gene; and it is hypothesized that imprinted genes have co-evolved with placenta and play an important role in placental/fetal development (Bischoff et al., 2009). Also, the PEG3 gene has been previously implicated with restricted fetal growth in mice models (L. Li et al., 1999). Building on this additional evidence of PEG3 association with growth restriction from mice models, Tsai et al performed an association study with 2 common SNPs in PEG3 (rs1055359 and rs33931963) and found a significant association of rs1055359 with SGA births in Caucasians (n=158) but not in African Americans (AA) (n=123) (Tsai, Hardison, Marks et al., 2010).

With added confirmation, the authors wanted to examine the functional role of additional variants in linkage disequilibrium with the associated SNP and also evaluate if the risk alleles comprised of common or rare alleles. It has been shown that rare variants can form a synthetic association with common variants and this can result in an association signal of the common marker with trait (Cirulli & Goldstein, 2010). For further evaluation, Tsai et al conducted a sequencing study of this gene from 55 placental samples (23 Caucasians and 32 AA). Although, no markers in the AA group were significantly associated with the SGA trait, the authors wanted to investigate the role of rare variants in the AA group with SGA.

The African ancestry has higher levels of genetic diversity than Caucasians, and it may be possible that the gene of interest is enriched with low frequency variants resulting in no association signal in the initial study. Such an approach has been successfully applied to implicate IL4 gene with asthma in African Americans (Haller, Torgerson, Ober, & Thompson, 2009). A 0.5 Mb region around and including PEG3 (Figure 4.1) was targeted using custom solution hybrid capture. All 55 samples were pooled together using unique barcodes and the pooled samples were sequenced over 2 lanes of Illumina Genome Analyzer IIx. All details of the capture can be found in (Tsai, Hardison, Keebler et al., 2010).

I analyzed the sequencing data for 55 de-multiplexed samples in this study. Thorough analysis was conducted for variant calling as well as association testing. I took advantage of the small dataset to test differences in pipeline setup for variant calling – specifically looked at three items: 1) alignment to entire human genome (hg19) vs. targeted chromosome; 2) the effect of downsampling on variant calling; and 3) the effect of duplicate read removal on variant calling. Single Nucleotide Polymorphisms (SNPs) and Insertion/Deletion (Indel) variations were called for all the samples and different checks were employed to test the integrity of called variants. Association testing was conducted with dichotomous trait of SGA and control. I analyzed variants in terms of frequency and putative functional effects with Polyphen-2 (Adzhubei et al., 2010); and used this information to conduct association testing between an accumulation of rare/functional variants in cases as compared to controls. No low frequency variants were found to be statistically significantly aggregated in SGA samples vs. controls.

Methods

Variant Calling:

The pipeline set up for the family data study (chapter 2) was adapted to call variants in this dataset. The pipeline was setup in bash scripts and run on a cluster of 30 commodity servers with 4 to 24 CPUs per node, 8 to 60 GB of RAM, and Ubuntu Linux as the operating system. FASTQ files (a standard format to store nucleotide sequence and base qualities) from the two

lanes for a given sample were merged into one FASTQ and the pipeline was run on one merged FASTQ. This was done for purposes of data management and because no differences in FASTQ qualities were observed between the two lanes (Figure 4.2). Sample 801 stands out in this plot because it showed a high proportion of duplicated reads (~55%). The following steps were performed for variant calling:

1. Barcode uniformity: A barcode serves the purpose of tracing fragmented DNA molecules back to their originating sample. All fragmented DNA molecules from one individual are ligated with the same unique barcode. After ligation with barcodes, the targeted DNA fragments from across the samples are pooled together and simultaneously sequenced on a single lane. Barcode uniformity looks at how evenly the number of reads from sequencing could be uniquely assigned to each sample. The FASTQ files in this study were already demultiplexed, i.e., reads were already assigned to each sample by Tsai et al (Tsai, Hardison, Keebler et al., 2010). We assessed the number of reads that could be assigned to each sample.
2. Adapter clipping and Low Quality Base Trimming: Fastq-mcf software (Aronesty, 2011) was used for adapter clipper and base trimming. Low quality bases towards the end of the reads were detected and clipped. Bases not called by Illumina (N's) were removed from the ends of the reads. Additionally, any adapter sequences were removed.
3. Base Quality Score Recalibration: A heuristic polynomial model was fit to correct the inflated Illumina scores as recommended in (DePristo et al., 2011).
4. Alignment of reads: The reads were aligned to the entire human genome (hg19) using BWA v0.5.9 (H. Li & Durbin, 2009). Default BWA parameters were used to generate sequence alignment (SAM) files and binary equivalent of SAM files (BAM). We also compared this to alignment of reads from five random samples to chr19 only. Statistics for SAM files were generated to assess alignment differences.
5. Realignment of BAM files: Genome Analysis Tool Kit version 1.4-30-gf2ef8d1 (GATK) was used to realign reads around insertions and deletions (indel) to minimize

spurious variant calls that may arise from mismatch SNPs. A secondary realignment was performed around dbSNP indels to call known indels in the 55 samples.

6. Variant Calling: GATK was used for preliminary variant calling (SNPs and indels) from the BAM files in reference to hg19 genome. No downsampling of reads was done. Variant calls were filtered as per the following criteria:
    a. Quality by Depth: QD<=2.5; This is a GATK parameter which is computed as phred scaled probability of observing a variant at the given site/depth at the site Low QD values may indicate errors (Broad Institute, 2011)
    b. Strand Bias: FS>=60; GATK parameter where high values indicate that mostly one of the strands is showing evidence for the variant.
    c. Mapping quality: MQ<55; GATK parameter that is a summary statistic (Root Mean Square) of the mapping qualities of all reads spanning a particular variant site. Low values are indicative of reads with low mapping quality.
    d. Depth: DP <= 5; Any reads with depth less than or equal to 5 were filtered
    e. If a known indel was called at a SNP locus after secondary realignment, the variant call was removed from the sample
    f. If a variant call was in a homopolymer run (HRun >= 5), had multiple allelic calls across samples of same ethnicity, or had other SNP calls in 20bp vicinity; then it was filtered
7. Duplicate read removal: One of the library preparation steps for Illumina is to amplify fragmented DNA molecules with polymerase chain reaction (PCR) technique. This step can sometimes result in amplification errors which may magnify during the PCR cycles and can be erroneously called as variants. Duplicate read removal removes any duplicate reads so that evidence from the same DNA fragment is not counted more than once. Duplicate read removal was done with Picard tools (http://picard.sourceforge.net) where all read pairs with identical 5' coordinates and orientation are marked as duplicates and removed (except for the best pair defined by the highest sum of base qualities). This step was only performed for randomly selected five samples as well as sample 801 (which showed highest percent of read

duplication) to assess differences in variant calling with and without duplicate read removal.

8. Downsampling coverage for variant calling: Depth of coverage (DoC) refers to how many reads cover a given base position. The GATK by default downsamples coverage for speed-up reasons in data processing of single samples. We wanted to evaluate the consequences of downsampling on variant calling and we did this by reducing the DoC to the default of 250 reads for variant calling by GATK for 5 random samples.
9. Quality Assessment: Variant quality assessment was done in terms of percent reads providing evidence for the variant allele, if the variant calls could separate the samples of the two ethnicities (via Principal Component Analysis) and dbSNP membership.
10. Putative Functional Evaluation: ANNOVAR and Polyphen-2 were used to annotate and evaluate putative effects of variants on protein function. Frequency of variant alleles was assessed in the sample populations.

Haplotype phasing:

PEG3 is an imprinted gene and only expression from paternal genome is observed in the placenta. (Tsai, Hardison, Marks et al., 2010). The SNPs in PEG3 region (chr19:57321445-57352094) were phased using BEAGLE (v3.3.2) (Browning & Browning, 2007) and SNPs from the paternal haplotype were used for association testing assuming that PEG3 gene expression is only influenced by SNPs on paternal haplotype.

Paternal haplotype was determined using allelic information for expressed haplotype; available through cDNA genotyping of 3 markers (rs1055359, rs3319363, rs34051133) (Tsai, Hardison, Marks et al., 2010). In some cases, where the fetus carried homozygous copies of the expressed allele at all three positions, it was not possible to differentiate the paternal haplotype from the maternal haplotype and these samples were dropped for association testing. We were left with 19 Caucasian and 20 AA samples.

Association testing with trait status (SGA vs. control):

Association testing was done separately for each ethnic group separately because no association signal was originally found in the AA group. We did not want to dilute the signal for significant association by mixing the two groups. Also, we were concerned that population stratification might falsely result in a significant association (Cardon & Bell, 2001).

1. Fisher's exact test was conducted for all SNPs found in the PEG3 region. Permutation testing was used to account for multiple hypothesis testing. This was done in BEAGLE (v3.2.2).
2. Weighted Sums test (Madsen & Browning, 2009) was modified to weight the variants by putative functional effect as predicted by Polyphen-2. Here, we tested to see if cases had more damaging variants than controls. Briefly, a genetic load was calculated for every individual based on the total number of variants it carried in the exonic regions of PEG3 and the probability score that the variant is damaging to protein function (synonymous SNPs had a score of 0 and a very damaging mutation had a score close to 1). A Wilcoxin rank sum test was done to compute the p-value for the alternate hypothesis that the difference is genetic loads between cases and controls is non-zero. The genetic load is computed across all marker alleles and hence multiple hypothesis correction is not necessary.
3. Weighted Sums test was implemented as suggested in (Madsen & Browning, 2009) for weighting variants as per their frequency (Steps A and B). First, the frequency of allele was computed in the control groups (p= probability of observing the allele in controls). Second, weight for each allele was computed as the standard deviation of binomial distribution [sqrt(n*p*(1-p))] where n is the total number of individuals in the study. As a result, alleles with p close to 0 and 1 in controls have a very high weight as compared to p=0.5. Third, genetic load was computed for every individual based on the number of variants and weights of each variant. Fourth, P-value was determined from a Wilcoxin Rank Sum Test (instead of a permutation analysis as originally implemented in the weighted sums test).

Results

Barcode Uniformity:

The two lanes (lane 5 and 6) of GAIIx resulted in 97.6M and 99.2M paired end reads respectively. About 4.66% reads from lane 5 and 4.82% reads from lane 6 were not assigned to any sample (Figure 4.3). Non-uniformity in read distribution across samples was observed; but this was consistent across the two lanes – i.e., the same samples got assigned the least and most number of reads in both lanes. It is difficult to establish if this is a result of barcode bias or certain samples were difficult to sequence than others.

Read Alignment:

Alignment with default BWA parameters to the entire genome aligned 99.56% on average (sd: 0.06) with median mapping quality of60. On average, 61.68% of the total reads aligned to chr19 (sd: 3.52). When we aligned only to chr19 and compared this to whole genome alignment, a mean increment of 1.63% (sd: 0.24) in read alignment to chr19 was observed for the five samples.

Variant Calling:

Variant filtering resulted in throwing out ~27% of SNPs and 68% of indels per individual. The filtering process removed a number of calls that had evidence from only a small proportion of reads (Figure 4.4). In the final set we were left with an average of 458 variants per individual (sd: 88) in the 0.5 Mb targeted region. The standard deviation in number of SNPs called per individual has no relation with the number of reads in sample, coverage of region, depth of coverage, or uniformity of coverage (data not shown). This may indicate that the differences in number of variants must be a characteristic of the sample. AA samples had more SNP calls than the Caucasians; 97.57% of calls made in Caucasians were also made in AA (Figure 4.5). Principal Component Analysis (PCA) was done using the final set of SNPs in all samples to check if the two ethnic groups could be separated dependent on their allelic composition (Figure 4.6). PCA tries to account for most variability using the least number of different attributes provided (allelic calls in this case). In figure 4.6, we can observe that

about 10% of variability between AA and CAU samples can accounted for using a subset of the allelic information provided. Thus, the two groups can be stratified using this SNP information. We observe clustering of some samples from AA group with the Caucasian samples, and one of the Caucasian samples is grouped with the AA. It is unclear if this is because of sample mislabeling; or these are self-reported ancestry and hence may not be accurate.

Effect on alignment to chromosome 19 on variant calling:

As mentioned earlier, alignment to targeted region increased the percent of reads aligned to chr19. We also made more variant calls after aligning to chr19 only – on average 22 (± 4) additional calls were made. This can be contributed to two factors: a) Reads that would have aligned to more than one region in the genome, could now unambiguously align to only one location; b) Mapping qualities of aligned reads improved as alignments became unique.

Effect of duplicate read removal:

No dramatic differences in variant calling were observed after removing duplicate reads for the randomly five selected samples. For the sample 801, that showed a high rate of duplication – 12 variants not called originally were made after duplication removal. This was a result of removal of reads containing indels and hence showing a slight overall mapping quality improvement to get through the stringent filters. There were 25 variant calls that were not made after duplicate read removal – some of this can be attributed to lowered depths.

Effect of downsampling:

DoC for each sample was reduced to 250 for variant calling. Downsampling did not have any striking effect on variant calling. This can be accounted to two reasons: 1) the downsampling factor was small given the original DoC per sample. 2) As can be seen in Figure 4.4, most heterozygous calls were coming from proportionally balanced reads (for variant and reference alleles); and hence the subsampling of reads must have worked well.

Variant call summary:

Tables 4.1 and 4.2 summarize the distribution of variants called in the PEG3 region for the paternal haplotype. The number of calls made per sample is divided by UTR3, UTR5, exonic (#non-synonymous calls in exons), intronic and non-coding RNA (ncRNA) SNPs (annotated with ANNOVAR (Wang, Li, & Hakonarson, 2010)). The ncRNA SNPs arise from the PEG3 antisense RNA 1 contained within the coordinates of PEG3 (Figure 4.1). A majority of the SNP calls are intronic (none with annotation) in splice sites as annotated from variant tables in Ensembl Genome Browser http://useast.ensembl.org/index.html). Two dbSNP indels (rs34172934, rs150275552) were found in the PEG3 region. rs34172934 is an insertion of A in the UTR3 region found at a frequency 10/19 in CAU and7/20 in AA. rs150275552 is an insertion of A in the intronic region and was only present in one sample. Figure 4.7 shows a distribution of all variants called in the 0.5 Mb region around PEG3 in the AA and CAU samples. It can be observed that most calls are intergenic, followed by intronic calls. This is not surprising as most of the region sequenced was interegenic and we would also expect these regions to be less conserved than exonic portions of genome.

Haplotype phasing:

Paternal and maternal haplotype determination was done using cDNA allelic information. For the 23 CAU samples, haplotypes for 19 could be unambiguously segregated into maternal and paternal. The remaining 4 samples were homozygous for the three genotyped alleles and had to be removed from association analysis. We were left with 12 controls and 7 cases with paternal haplotypes. For the AA samples, 6 samples had no expressed allele genotypes available and an additional 6 samples, the paternal haplotype could not be unambiguously identified. We were left with 20 samples – 5 cases and 15 controls.

Association Testing:

A SNP-by-SNP association test (without indels) was done in BEAGLE using Fisher's exact test and a permutation-based multiple hypothesis corrected p-values were used to assess significance at α=0.05. The CAU population had two significant hits (rs1055395 and

rs9304785) (corrected p-val: 0.04096). Table 4.3 shows presents a list of top hits. The AA population showed no significant hits.

The exonic SNPs (synonymous and non-synonymous) were used for low frequency variant analysis with Weighted Sums Test (WST). We only tested for exonic variants with weights calculated via predicted functional score; because these variants are most relevant to protein function. For the CAU samples, 7 exonic SNPs were used with frequency ranging from 0.05 to 0.32 in the 19 haploid samples. The AA samples had 12 exonic SNPs included in the test with frequency range of 0.05 to 0.20 in the 20 haplotypes. No significant results were observed for the AA samples or CAU samples using either of the two WSTs; i.e; I could not conclusively say that the paternal haplotypes in SGA are enriched for low frequency exonic variants than non-SGA samples. It is should be noted that 2/5 AA SGA vs. 4/15 AA controls, and 4/7 CAU SGA vs. 0/12 CAU controls carried no exonic variants, which explains the observed statistical result.

I tested the intronic SNPs with WST using allele frequency with weights. We observed significance accumulation of intronic variants in CAU controls vs. cases (p-value = 0.0023). No significant differences were observed for AA cases vs. controls.

Discussion and Conclusions

This study was designed by Tsai et al. as a follow up for a previously discovered association signal on the paternally expressed PEG3 gene (rs1055359) with Small for Gestational Age (SGA). A subset of samples used in the original study was selected for multiplexed deep sequencing of PEG3 and the region around it to fine map the association signal. The question under consideration was if SGA samples were enriched for low frequency variants on the PEG3 expressed haplotype as compared to controls.

I modified the pipeline in Chapter 2 to test if the reads should be aligned to only the targeted region or to the entire human genome. I conclude that even though targeted alignment results in a slightly higher mapping of reads, this is not a good strategy as these reads do not necessarily map uniquely to the targeted region and can inflate the number of false positive calls. I also examined the effect of duplicate read removal especially since one

of the samples had very high amount of duplicated reads. Since I did not see a drastic difference in variant calling for this particular sample, I concluded that erroneous variant calling through PCR duplication was not a concern for other samples that showed substantially lower read duplication. Also since this was a small dataset, I decided to look at the effect of default downsampling vs. no downsampling on variant calling. I did not observe any dramatic differences and decided to use all read data for variant calling. I also modified the variant filtering strategy from Chapter 2 to make it more stringent as I did not have family data to check for consistency in variant calls.

Through deep sequencing, I discovered 58 PEG3 SNPs and 1 indel in 19 Caucasians; and 124 SNPs and 2 indels in 20 AA individuals. Of these 7 exonic SNPs in Caucasians and 12 exonic SNPs in AA were used for low-frequency association testing. No exonic variants were found to be significant hits. I did find a significant accumulation of intronic variants in CAU controls as compared to cases (p-value=0.0023). One has to be careful about interpreting this result because none of the intronic SNPs that had annotation available (via Ensembl) were predicted as splicing sites; and non-splicing intronic regions maybe less conserved in our genomes. For the AA samples, no tests were significant.

A difficulty that arises in making meaningful inferences for these results is that SGA is a complex trait, it is difficult to characterize, and is influenced by a lot of external factors (Cnattingius, 1997; Monk & Moore, 2004; Nohr et al., 2009). The power of the study depends on how accurately the samples were classified and accounting for covariates like smoking status, parity, weight of the mother (which was not done in this analysis). Also, PEG3 is an imprinted gene – the gene is epigenetically modified to control expression from one of the parental genomes. Epigenetics may also be influencing the expression levels of this gene. This study was also limited by small sample size (after accounting for the two ethnicities). Of note, the original association is present in the anti-sense transcript of PEG3 and is shown to be evolutionarily conserved (Choo, Kim, & Kim, 2008). It may be possible that PEG3 expression is controlled by PEG3-AS1 and this is the only SNP of significant association in Caucasian population. This will need to be followed up by association testing of rs1055359 with PEG3 gene expression levels. In conclusion, no significant association

was detected with low frequency exonic alleles and SGA; leading to suggest that the original signal maybe the only source of explanation for association.

Acknowledgments

I would like to thank Expression Analysis Inc, Durham, NC for allowing us to use their cluster to process this dataset.

Table 4.1: Variant calls in PEG3 for African American samples. This table gives a breakdown of the number of variant alleles called for the AA paternal haplotypes in the PEG3 region (chr19: 57321445-57352094) as per their SNP class. This region contained 124 unique markers. Number of markers in dbSNP is also indicated in the total SNPs column.

Sample	UTR3	ncRNA	Exonic (non-synonymous)	intronic	UTR5	Total (dbSNP)	SGA
3	4	1	0	12	0	17 (17)	Y
22	1	0	2 (0)	6	0	9 (9)	Y
617	1	0	1 (0)	10	0	12 (11)	
646	1	2	1 (0)	14	0	18 (18)	
659	3	1	1 (0)	16	1	22 (22)	Y
694	0	0	0	1	0	1 (1)	Y
710	0	1	2 (0)	15	0	18 (18)	
726	3	1	0	8	0	12 (12)	
734	5	1	1 (0)	9	0	16 (15)	Y
743	0	1	2 (0)	14	0	17 (17)	
756	0	1	1 (0)	14	1	17 (17)	
766	1	0	1 (0)	2	0	4 (4)	
767	3	1	1 (1)	13	0	18 (18)	
775	4	1	2 (1)	16	0	23 (23)	
798	3	1	1 (1)	11	0	16 (16)	
623	3	1	0	17	0	21 (20)	
624	0	0	0	1	0	1 (1)	
657	4	1	1 (1)	9	0	15 (15)	
668	3	1	0	7	0	11 (11)	
705	0	0	3 (2)	4	0	7 (7)	

Table 4.2: Variant calls in PEG3 for Caucasian samples. This table gives a breakdown of the number of variant alleles called for the CAU paternal haplotypes in the PEG3 region (chr19: 57321445-57352094) as per their SNP class. The region comprised of 58 unique markers. Number of markers with dbSNP membership is indicated in the total SNPs column.

Sample	UTR3	ncRNA	Exonic (non-synonymous)	intronic	UTR5	Total (dbSNP)	SGA
16	0	0	0	0	0	0	Y
34	3	1	1 (1)	14	0	19 (18)	Y
626	3	1	1 (1)	13	1	19 (18)	
652	5	1	1 (0)	18	1	26 (25)	
662	5	1	1 (0)	5	0	12 (12)	
674	0	1	1 (0)	15	0	17 (16)	
678	4	1	1 (0)	19	1	26 (25)	
682	5	1	1 (0)	17	1	25 (24)	
693	0	0	0	1	0	1 (1)	Y
698	0	0	1 (1)	1	0	2 (2)	Y
730	0	1	2 (1)	15	1	19 (18)	
738	5	1	1 (0)	5	1	13 (13)	Y
740	3	1	0	13	0	17 (16)	Y
746	5	1	1 (0)	16	1	24 (23)	
757	0	0	0	10	0	10 (9)	Y
761	1	1	1 (0)	13	0	16 (14)	
763	0	1	1 (0)	12	1	15 (14)	
764	0	1	2 (1)	14	0	17 (16)	
787	4	1	3 (1)	17	0	25 (24)	

Table 4.3: Top hits for SNP association with SGA. Top hits for SNP association with SGA trait in the 19 CAU samples are shown.

Chromosome location	dbSNP ID	Fisher's exact test pvalue	Corrected pvalue	SNP class
chr19 57324340	rs1055359	0.009030	0.04096	ncRNA
chr19 57331912	rs9304785	0.009030	0.04096	Intronic
chr19 57335179	rs1860566	0.03793	0.1788	Intronic
chr19 57339207	rs4801389	0.03793	0.1788	Intronic
chr19 57341067	rs1860567	0.03793	0.1788	Intronic
chr19 57344913	rs2870477	0.03793	0.1788	Intronic
chr19 57344967	rs2159551	0.03793	0.1788	Intronic
chr19 57351306	rs7258812	0.03793	0.1788	Intronic

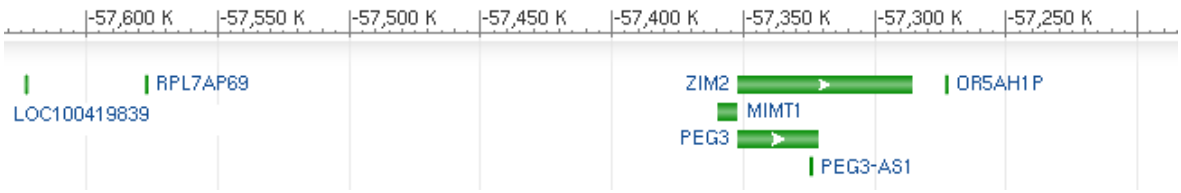

Figure 4.1: Genes in targeted region. The genes contained in the targeted region chr19:57,184,288-57,630,792 are shown. There are 2 known genes (ZIM2, PEG3), a known non-coding RNA (PEG3-AS1), a predicted gene (MIMT1) and 3 pseudo genes (RPL7AP69, OR5AH1P, LOC100419839) in the targeted region. This figure has been adapted from NCBI Entrez diagram for this region of interest (http://www.ncbi.nlm.nih.gov/gene/5178).

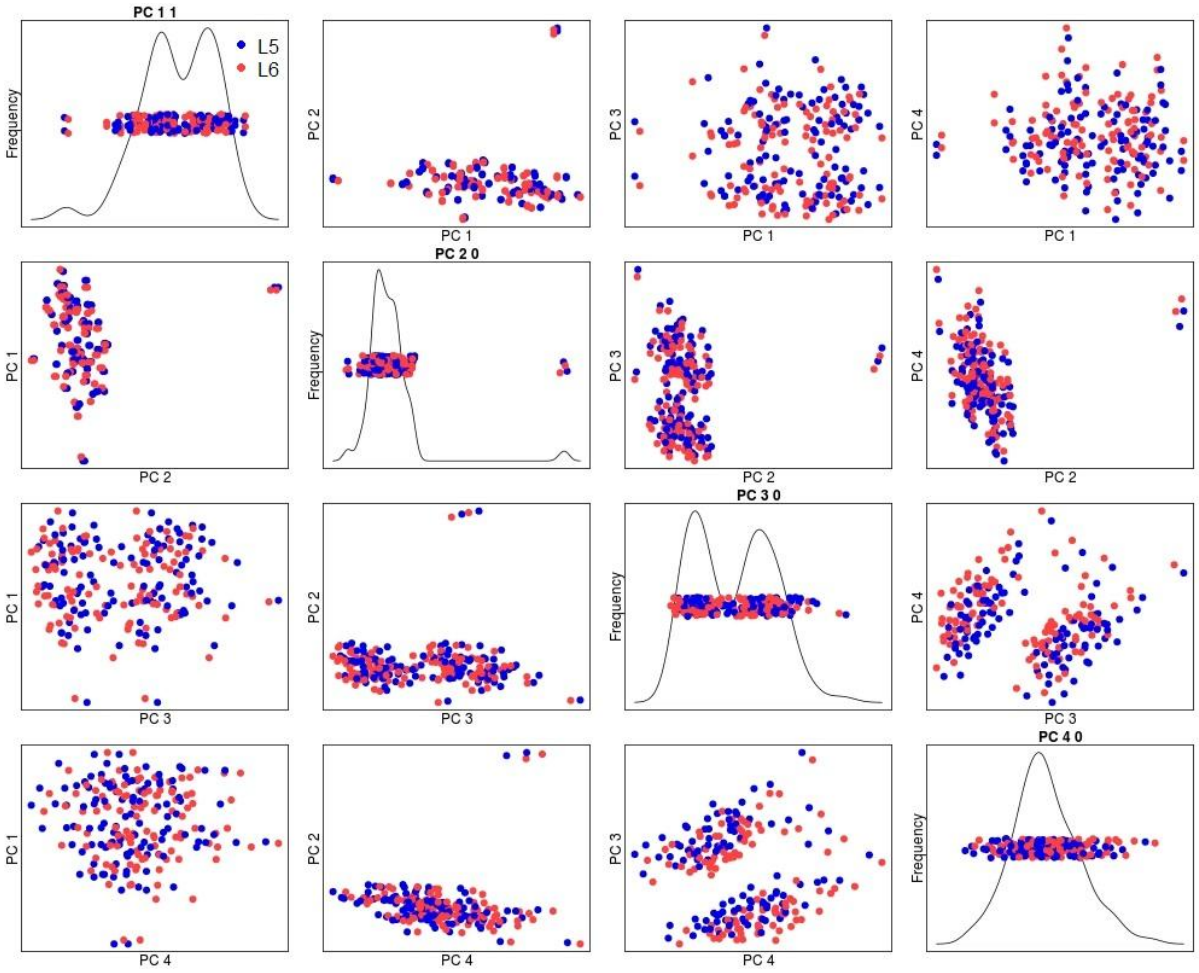

Figure 4.2: Principal Components Analysis on FASTQ qualities. Principal Components Analysis (PCA) was performed on fastq-qualities for the sample replicates run on lanes 5 and 6 of the flowcell. PCA was provided with multiple metrics like #reads per sample,, %duplicate reads, mean #duplicates, s.d. #duplicates, %A, %C, %G, %T, %N, min base quality, max base quality, mean base quality, and s.d. base quality. Using minimum number of these metrics, PCA tries to account for maximum variability in the original data. The two lanes are visually clustered together which leads to the conclusion that no quality differences were observed for the sample replicates between two lanes. As a result, they were merged into one fastq for running through analysis pipeline. The points that stand out in the first PCA component belong to sample 801 which showed a high level of duplicated reads in both the lanes.

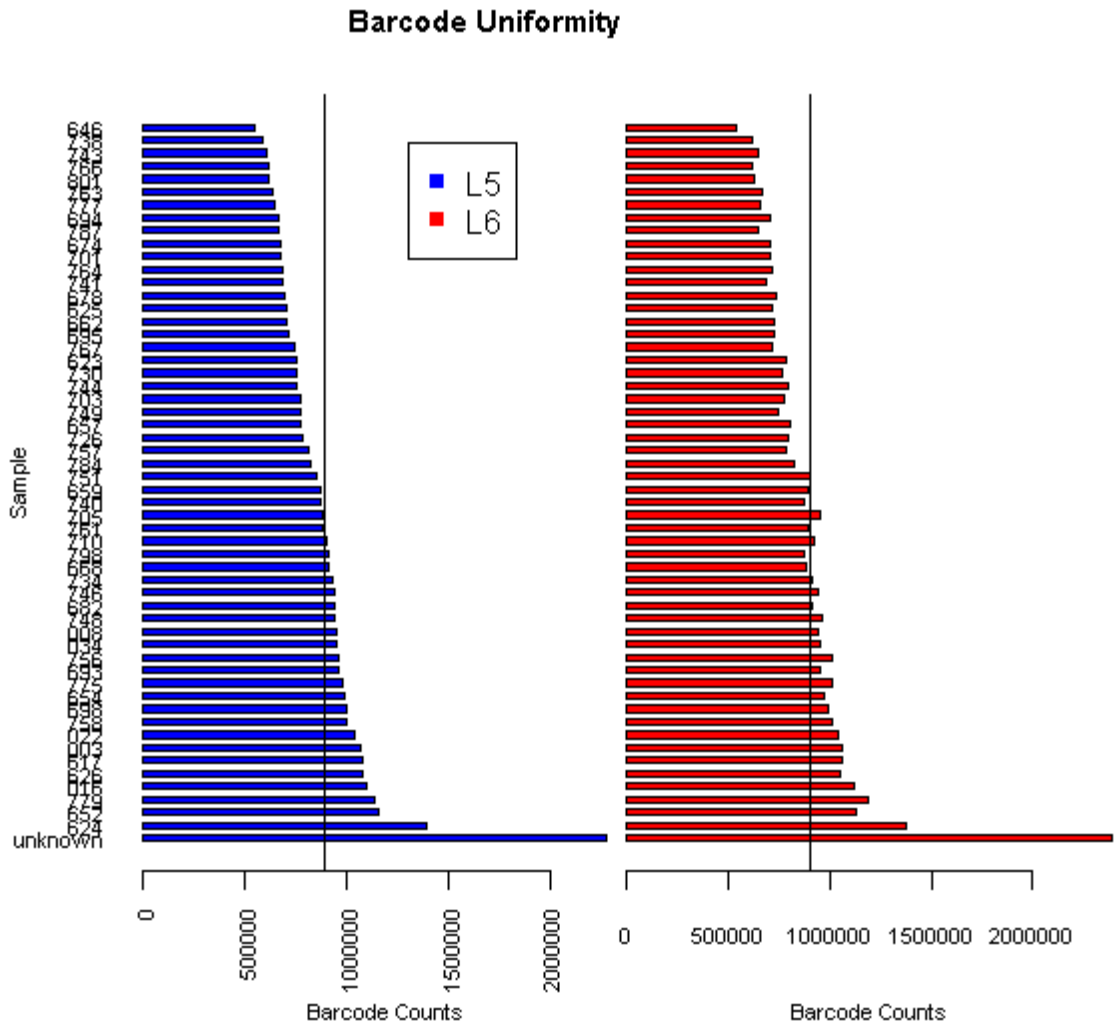

Figure 4.3: Barcode uniformity for multiplexed samples. Barcode uniformity for the 55 multiplexed samples on the two lanes is shown. The vertical line indicates the expected barcode counts per sample (total reads generated in lane/55). Non-uniformity in barcode performance is observed but the bias is consistent across the two lanes.

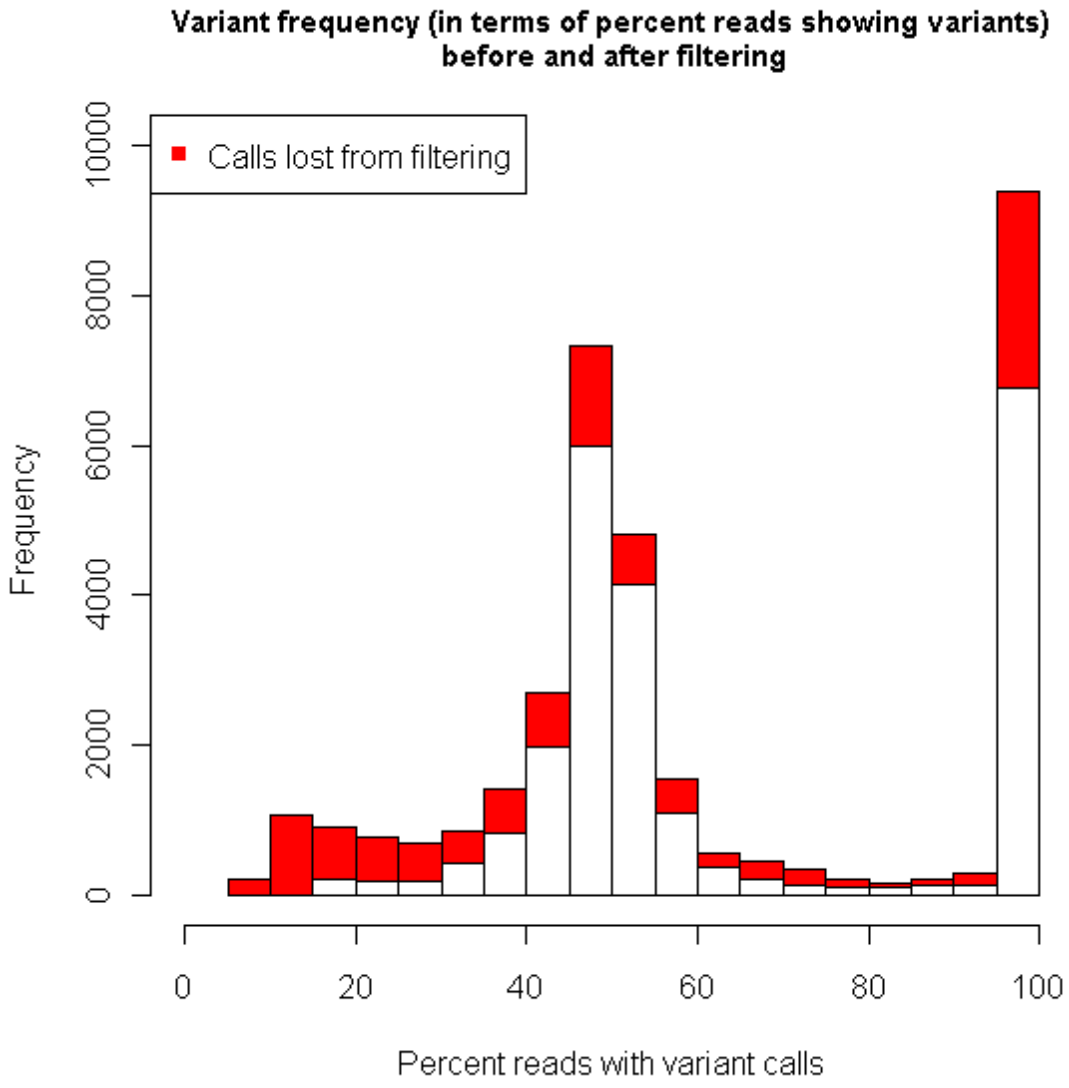

Figure 4.4: Proportion of reads supporting filtered calls. This figure shows a histogram of what proportion of reads at a given position provided evidence for the non-reference allele. The red area indicates the number of variants filtered out. Almost 200 variants were called when only 5-9% of reads at a given position contained alternate allele – all these variants were filtered out using different filtering criteria. (Please note that percent reads with variant calls was not used as a filtering criteria – it is simply used to check what category of variants were filtered).

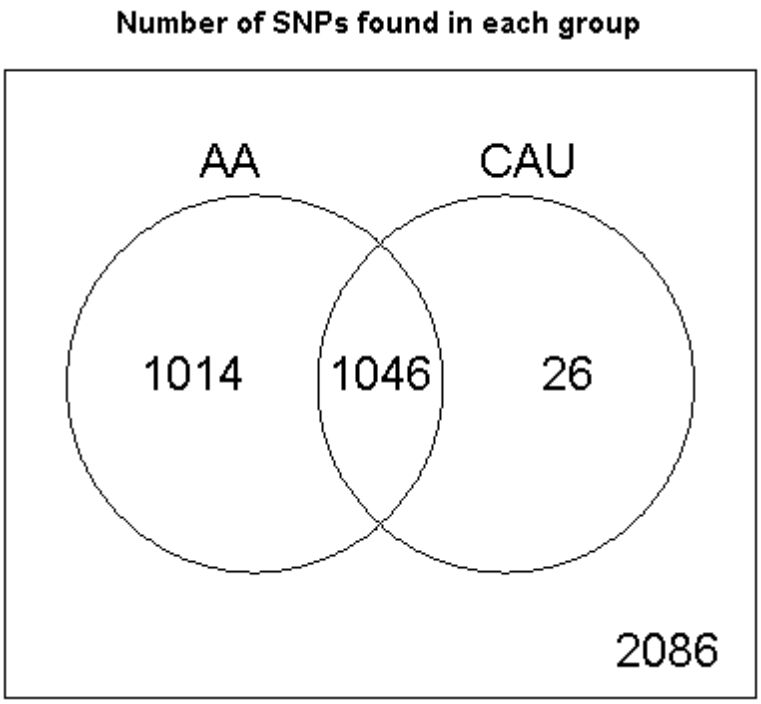

Figure 4.5: Distribution of SNPs in Caucasians and African Americans (AA). A total of 2086 positions were called as SNPs in the 0.5Mb targeted region. Of these 2086 SNPs, 1046 are shared between Caucasians and AA, 1014 are exclusive to AA, and 26 are exclusive to Caucasians.

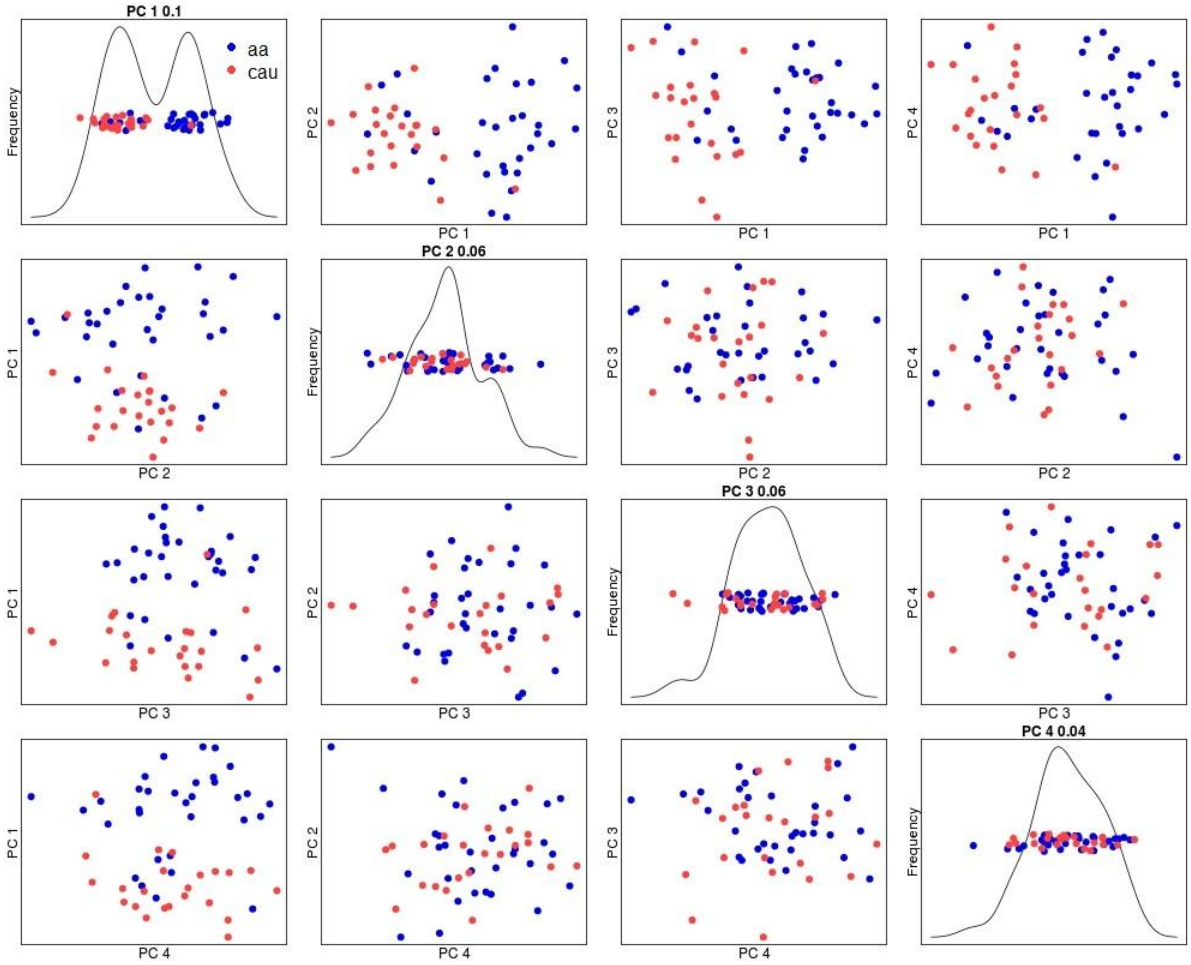

Figure 4.6: Stratification of ethnic samples. Principal Component Analysis (PCA) was done on the final SNP set (2086 SNPs) to check if the two ethnic groups can be separated. PCA tries to account for most variability using the least number of different attributes provided (allelic calls in this case). In this figure we can observe that about 10% of variability between AA and CAU samples can accounted for using a subset of the allelic information provided.

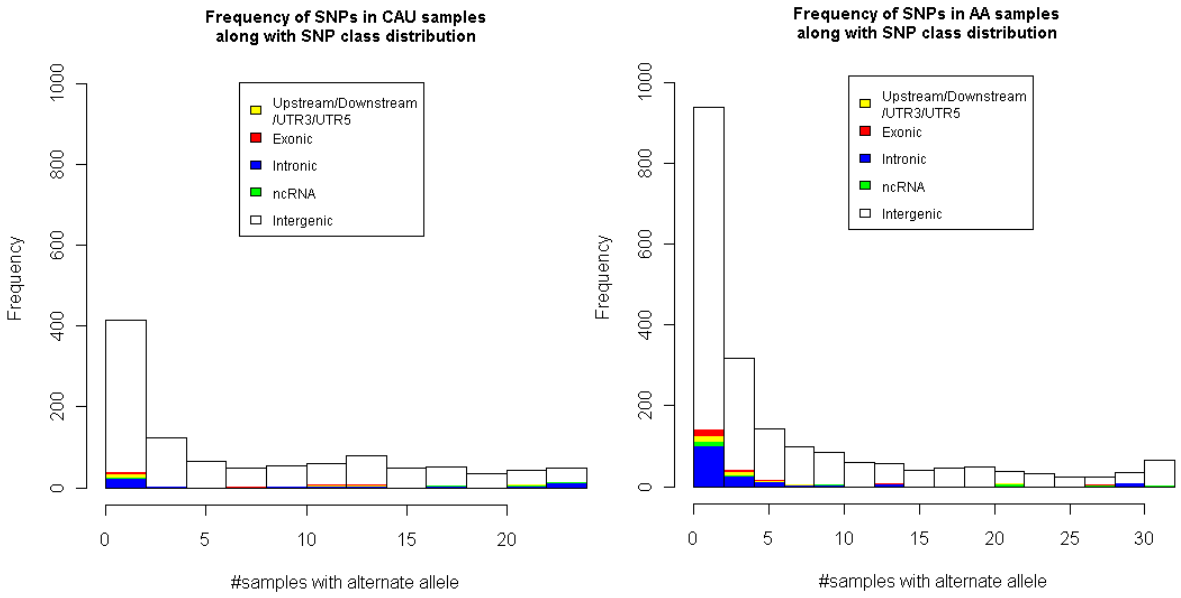

Figure 4.7: SNP distribution by class. SNP distribution by class (UTR3, UTR5, Exonic, Intronic, ncRNA, Intergenic) is shown for all variants called in AA and CAU samples. Most calls were intergenic and present in low frequency in these samples.

References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. *Nature Methods, 7*(4), 248-249. doi:10.1038/nmeth0410-248

Aronesty, E. (2011). *Command-line tools for processing biological sequencing data*

Bischoff, S. R., Tsai, S., Hardison, N., Motsinger-Reif, A. A., Freking, B. A., & Piedrahita, J. A. (2009). Functional genomic approaches for the study of fetal/placental development in swine with special emphasis on imprinted genes. *Society of Reproduction and Fertility Supplement, 66*, 245-264.

Broad Institute. (2011). *Understanding the unified genotyper's VCF files.* Retrieved June 16, 2012, from http://www.broadinstitute.org/gsa/wiki/index.php/Understanding_the_Unified_Genotyper%27s_VCF_files

Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics, 81*(5), 1084-1097. doi:10.1086/521987

Cardon, L. R., & Bell, J. I. (2001). Association study designs for complex diseases. *Nature Reviews.Genetics, 2*(2), 91-99. doi:10.1038/35052543

CDC. (2008). *QuickStats: Percentage of small-for-gestational-age births, by race and hispanic ethnicity---united states, 2005* (Morbdity and Mortality Weekly Report No. Vol. 57 / No. 50). Washington, DC: U.S. Government Printing Office.

Choo, J. H., Kim, J. D., & Kim, J. (2008). Imprinting of an evolutionarily conserved antisense transcript gene APeg3. *Gene, 409*(1-2), 28-33. doi:10.1016/j.gene.2007.10.036

Cirulli, E. T., & Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews.Genetics, 11*(6), 415-425. doi:10.1038/nrg2779

Cnattingius, S. (1997). Maternal age modifies the effect of maternal smoking on intrauterine growth retardation but not on late fetal death and placental abruption. *American Journal of Epidemiology, 145*(4), 319-323.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics, 43*(5), 491-498. doi:10.1038/ng.806

Haller, G., Torgerson, D. G., Ober, C., & Thompson, E. E. (2009). Sequencing the IL4 locus in african americans implicates rare noncoding variants in asthma susceptibility. *The Journal of Allergy and Clinical Immunology, 124*(6), 1204-9.e9. doi:10.1016/j.jaci.2009.09.013

Kristensen, S., Salihu, H. M., Keith, L. G., Kirby, R. S., Fowler, K. B., & Pass, M. A. (2007). SGA subtypes and mortality risk among singleton births. *Early Human Development, 83*(2), 99-105. doi:10.1016/j.earlhumdev.2006.05.008

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics (Oxford, England), 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Li, L., Keverne, E. B., Aparicio, S. A., Ishino, F., Barton, S. C., & Surani, M. A. (1999). Regulation of maternal behavior and offspring growth by paternally expressed Peg3. *Science (New York, N.Y.), 284*(5412), 330-333.

Madsen, B. E., & Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics, 5*(2), e1000384. doi:10.1371/journal.pgen.1000384

Monk, D., & Moore, G. E. (2004). Intrauterine growth restriction--genetic causes and consequences. *Seminars in Fetal & Neonatal Medicine, 9*(5), 371-378. doi:10.1016/j.siny.2004.03.002

Nohr, E. A., Vaeth, M., Baker, J. L., Sorensen, T. I., Olsen, J., & Rasmussen, K. M. (2009). Pregnancy outcomes related to gestational weight gain in women defined by their body mass index, parity, height, and smoking status. *The American Journal of Clinical Nutrition, 90*(5), 1288-1294. doi:10.3945/ajcn.2009.27919

Tsai, S., Hardison, N., Keebler, J., James, A., Motsinger-Reif, A., Bischoff, S., et al. (2010). Targeted solution hybrid capture, barcoding, and multiplexed massively parallel sequencing of PEG3 0.5 mb genomic interval in 55 human individuals. (PhD, North Carolina State University).

Tsai, S., Hardison, N., Marks, O., James, H., Motsinger-Reif, A., Bischoff, S., et al. (2010). Association of paternally inherited variants of PEG3 with human SGA birth. (PhD, North Carolina State University).

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research, 38*(16), e164. doi:10.1093/nar/gkq603

CHAPTER 5

Concluding remarks

Gunjan D. Hariani

Bioinformatics Research Center, North Carolina State University, Raleigh NC 27695

Dissecting genetics of common complex diseases is a multifaceted problem. Multiple genes, gene products and regulatory factors work together to manifest a phenotype. No one gene is necessary or sufficient to result in a diseased phenotype. As a result, the hypothesis free approach of Genome Wide Association Studies (GWAS) is a much appealing strategy over candidate gene study for discovery purposes and has been widely applied to many genetic traits. But, GWAS relies on the principle of common disease common variant (CDCV) – as per this theory, the alleles responsible for common diseases will also be common in the population (MAF > 5%) with low to moderate effect sizes (Reich & Lander, 2001). This seems logical because one would expect high risk alleles that affect reproductive fitness to be less frequent in the population; whereas those alleles that cause late onset diseases maybe more prevalent as they can escape selection pressure. Thus, GWAS is only hypothesis free in the sense that it does not confine the association of variants to certain genes but the underlying principle of GWAS is restricted to only common variants. This may partly explain the limited success of GWAS to only those traits where this genetic architecture may actually hold true.

Researchers, especially in human genetics, are now shifting focus to examine low frequency variants due to large fraction of unexplained heritability of complex diseases by common variants (Manolio et al., 2009). The focus on low frequency variants also allows researchers to examine an alternate hypothesis – common disease rare variant (CDRV) (Pritchard, 2001) in contrast to the one suggested by the CDCV theory (Reich & Lander, 2001). As per this hypothesis, alleles with moderate effect sizes may be individually rare but collectively common in the population to result in diseased phenotype. DNA sequencing opens the avenue of exploring the genomic space beyond common variants (MAF > 5%) that

are investigated in GWAS. Even in cases where GWAS have successfully associated variants with traits, the variants themselves may not be causal and the functional variants may lie in linkage disequilibrium with the associated variant. Consequently, sequencing regions around associated GWAS signal holds promise to discover functional variants and additionally explain a portion of missing heritability (Manolio et al., 2009).

Advancement in sequencing technology is allowing researchers to explore the genomes on a finer scale and at unprecedented speed, volume and cost. The voluminous sequencing data comes with its own slew of challenges. Before we can proceed to analysis, we have to think about the infrastructure needed to store and process the volumes of data generated by these sequencers. Solutions like cloud computing provide a remedy this problem. The next big challenge is putting together the pipeline necessary to process this data. A lot of work has already been put into developing fast and efficient algorithms to handle these voluminous datasets. The difficulty arises in selecting the right software for an instrument specific data and to optimize the data usage by tweaking the software parameters. This ties in with understanding the underlying principle of an instrument, the systematic biases introduced by that instrument, and limitations of the software.

Processing sequencing data requires integration of various software packages, database resources, scripting, automation, and various checkpoints for preventing pipeline failure. There is no set standard protocol to do this (although the BROAD institute is moving one step closer towards this and setting up recommendations to put together a pipeline). But, given that there are numerous applications of sequencing, many different sequencing technologies, and a multitude of software packages; every researcher will have their own preference and bias towards a particular package. A biologist with limited computing background has to deal with setting up the hardware and software to analyze the data.

In this dissertation, I have touched on a few of the above mentioned issues as next generation sequencing (NGS) was the technology used to genotype samples over some regions of interest and identify variants in these samples. Chapter 2 summarizes the rationale of following the CDRV theory and addresses some of the statistical challenges that arise when working with rare variants due to high dimensionality of data. In this chapter, I have

conducted simulation studies with varying parameters like sample size, number of causative variants, effect size of variants, and addition of neutral variants; to assess the powers of three methods designed for testing the rare variant hypothesis. I conclude that for candidate gene studies and unidirectional effects of variants, these studies will be powerful to detect association with pooling of rare variants.

In Chapter 3, Illumina NGS technology has been used to detect different forms of variants in cell lines and carry out association tests with drug response for 30 FDA approved drugs. Here, sequencing was applied to discover variants that may better explain inter-individual variability to drug response. A lot of time was spent putting together the pipeline to call variants and conducting checks at every level of processing (raw FASTQ, processed FASTQ, post-alignment, targeted capture, and variant calling). The variant calling process was shell scripted in order to be able to easily replicate results with the same dataset in future if needed; or apply the same pipeline to a different dataset for a similar study to avoid biases from software choices. Different variant calling software packages were tried before settling on the Genome Analysis Tool Kit variant caller (McKenna et al., 2010) developed by the BROAD institute. After variant calling, variants were annotated and tested for genotyping accuracy using various sources like dbSNP (Sherry et al., 2001), HapMap (International HapMap Consortium, 2003), Annovar (Wang, Li, & Hakonarson, 2010), Polyphen (Adzhubei et al., 2010), and Ensembl (McLaren et al., 2010). These steps relate to the choice of software for analysis and integration of software packages for putting together a pipeline. Next, association testing was carried out and significant hits were found for two of the drugs used in the study. These results are very exciting and one of the next steps would be to follow up these results. This could be done either with knockdown experiments in cell lines or in a case – control association setting in an independent cohort. The other major steps would be to test for accumulation of rare variants in responders vs. non-responders and further investigate the role of rare variants in the candidate genes selected for this study.

In Chapter 4, the pipeline developed in chapter 3 was modified to a sequencing dataset designed to identify variants around a previously discovered association signal in samples from 2 populations – Caucasian and African American. There is sufficient evidence

in the literature that rare variants can better explain some of the signals identified through association of common variants and can further elucidate the genetic architecture of a trait. This understanding was applied to a pregnancy related trait – small for gestational age (SGA); identified rare variants and implemented one of the methods that had relatively high power in simulation studies from Chapter 2 to test for excess accumulation of rare variants in cases vs. controls. I concluded that infants with SGA are not enriched with functional rare variants in the gene of interest (PEG3) as compared to controls, in both Caucasian and African American samples.

Looking ahead in the avenue of low frequency variants, the success of this new theory will be contingent upon the proper design of association studies; clear definitions of phenotype; along with development of apt statistical tools that tackle the high dimensionality of data. Common diseases have an added layer of phenotype complexity – multiple symptoms together define a diseased state. One tactic to deal with complex phenotype definitions could be to divide the trait of interest into different subtypes (e.g. based on the various symptoms observed) and carry out association studies with individual subtypes. Splitting a phenotype into different components may make the cases more similar and improve the power of association, especially in a scenario where we are testing large number of variants.

With regards to design of association studies, important lessons have been learnt from GWAS – like appropriate matching of case/control subjects, accurate and extensive phenotype measurements, cohort sizes, and replication along with prioritization of results. These will be extended into studies testing for rare variant association with phenotypes where replication is going to be more complex than in the case of common variants. If variants are individually rare, validation techniques will have to account for missing out replication due to rarity of a variant rather than categorizing it as a false positive finding; and also for sifting biologically meaningful variants from background rare variants. New rules will have to be set up to validate the findings of initial studies. I think that preliminary studies should combine evidence from other biological sources before trying to re-establish the initial association in a separate cohort. Discovery of disease causative alleles is a time consuming, expensive, and

tedious process; and a researcher would want to invest money and time in pursuing probable high risk and biologically relevant variants. Adding a biological component in the association studies at an early discovery stage will help to narrow down potential variants for replication and improve power of association at subsequent stages. One of useful biological confirmations could be integration of intermediate RNA expression levels because the observed phenotype expression goes through multiple stages of control at DNA, RNA and protein levels. It is quite possible that differences in levels of RNA expression between cases and controls due to differences in genetic variants may better explain phenotypic variance. It would be worthwhile to examine if an associated variant is an eQTL – i.e., it regulates RNA expression. Pleiotropy of genes could lead to another source of validation – we know that some genes are responsible for controlling multiple phenotypes. If a gene known to regulate multiple traits is associated in the initial study, then replication studies could measure some additional phenotypes (conditional to low cost, ease of measurement, and low ambiguity in phenotype classification) and test for association of the initial findings with these phenotypes. During the replication stages, one other approach could estimate a biological effect size by performing pathway analysis and determining if genes with deleterious mutations can be rescued by other gene products. Genes that cannot be rescued could be prioritized for further investigation or validation through maybe gene knockout studies.

Considering the two hypotheses mentioned here (CDCV and CDRV), there are scenarios where both these hypothesis may fail to account association of traits with variants due to various reasons. One of them could be due to low statistical power of the study – for example there could be scenarios where these hypotheses hold but where the effect size of each contributing factor is too low, then the association studies may not have sufficient power to establish a relation between phenotype and genetic variants. These two theories only account for main effects of variants and do not consider interaction effects between SNPs. Epistasis is another facet of complex diseases that will furthermore explain the observed variance in phenotypes. Additionally, variants other than SNPs like copy number variants, genomic deletions or insertions, duplications and inversions also contribute to phenotype variance. Incorporation of variants other than SNPs in discovery phase of

association studies may accelerate findings of genes that contribute significantly to an observed phenotype.

The shift in focus from common variants to low frequency variants does not imply that one expects these to explain the remaining unexplained heritability for all traits. I believe that technology is a driving force in determining how science proceeds. Today, the sequencing technology is allowing us to pick off from where GWAS left and associate the remainder of our genomic variants with phenotypic traits in a cost-effective manner. Sequencing data is allowing us to test hypothesis that were previously not possible to investigate and examine the genetic architecture of a trait. If the hypothesis tested does not hold for a particular trait, then we know that there is more in the genome that is influencing the trait. It seems to me that our genomes are too complex to be sufficiently elucidated by just two contradicting theories; although the objective of testing these theories is not to ascertain the entire genetic architecture of traits. It is rather to discover variants of significance in clinical settings or other important human traits; and the missing heritability is serving as a measure to determine how well each theory holds for a particular trait.

References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. *Nature Methods, 7*(4), 248-249. doi:10.1038/nmeth0410-248

International HapMap Consortium. (2003). The international HapMap project. *Nature, 426*(6968), 789-796. doi:10.1038/nature02168

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature, 461*(7265), 747-753. doi:10.1038/nature08494

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research, 20*(9), 1297-1303. doi:10.1101/gr.107524.110

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., & Cunningham, F. (2010). Deriving the consequences of genomic variants with the ensembl API and SNP effect predictor. *Bioinformatics (Oxford, England), 26*(16), 2069-2070. doi:10.1093/bioinformatics/btq330

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *American Journal of Human Genetics, 69*(1), 124-137. doi:10.1086/321272

Reich, D. E., & Lander, E. S. (2001). On the allelic spectrum of human disease. *Trends in Genetics : TIG, 17*(9), 502-510.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbSNP: The NCBI database of genetic variation. *Nucleic Acids Research, 29*(1), 308-311.

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research, 38*(16), e164. doi:10.1093/nar/gkq603

APPENDIX

Appendix A – Supplementary tables and figures from Chapter 3

Table S.1: Pedigree structure for 95 CEPH LCLs. The 95 CEPH LCLs from 14 families used in the study are listed in this table along with the pedigree structure.

Family	Sample	Relation
35	12615	father
35	12616	mother
35	12617	son
35	12618	son
35	12619	daughter
35	12620	daughter
35	12621	daughter
35	12622	daughter
35	12623	son
45	12698	father
45	12699	mother
45	12700	son
45	12702	son
45	12703	son
45	12704	daughter
45	12705	son
45	12706	son
45	12849	daughter
1334	10846	father
1334	10847	mother
1334	12138	son
1334	12139	daughter
1334	12141	son
1334	12142	son
1340	7019	mother
1340	7027	son
1340	7029	father
1340	7040	son
1340	7062	daughter
1340	7342	son

Table S.1 continued

1340	11821	Son
1341	6991	mother
1341	7006	daughter
1341	7012	daughter
1341	7020	son
1341	7021	son
1341	7044	daughter
1341	7048	father
1341	7343	daughter
1345	7345	maternal grandmother
1345	7348	mother
1345	7357	maternal grandfather
1350	10855	mother
1350	10856	father
1350	11822	son
1350	11824	daughter
1350	11825	daughter
1350	11827	daughter
1362	10860	father
1362	10861	mother
1362	11983	daughter
1362	11984	son
1362	11985	daughter
1362	11986	daughter
1362	11988	daughter
1362	11989	daughter
1408	10830	father
1408	10831	mother
1408	12147	daughter
1408	12148	son
1408	12149	daughter
1408	12150	daughter
1408	12151	daughter
1408	12157	daughter
1420	10838	father
1420	10839	mother

Table S.1 continued

1420	11999	Daughter
1420	12001	daughter
1420	12002	daughter
1420	12007	son
1447	12752	father
1447	12753	mother
1447	12754	daughter
1447	12756	son
1447	12765	son
1451	12766	father
1451	12768	son
1451	12770	daughter
1451	12771	son
1451	12772	daughter
1451	12773	daughter
1451	12774	son
1451	12848	daughter
1454	12802	mother
1454	12803	daughter
1454	12805	son
1454	12806	son
1454	12807	daughter
1454	12810	son
1459	12864	father
1459	12866	son
1459	12868	daughter
1459	12869	daughter
1459	12870	son
1459	12871	son

Table S.2: List of 103 candidate genes sequenced in the study. The genes were selected based on their involvement in pathways for drug metabolism, transport, or drug action for 5 classes of chemotherapy drugs: fluoropyrimidines, anthracyclines, platinum compounds, taxanes, and camptothecins.

Gene Symbol			
MTHFR	POLH	MT1A	UGT2B15
AKR1A1	REV3L	CES2	UGT2B7
DPYD	PMS2	NQO1	ABCG2
GSTM1	POLM	TP53	NFKB1
UCK2	UPP1	TOP2A	DHFR
FASLG	ABCB1	MAPT	PSMC1
PARP1	CYP3A5	ABCC3	DUT
RRM2	CYP3A4	NME1	ERCC4
XDH	POLB	NME2	ABCC1
CYP1B1	GGH	MPO	ABCC6
MSH2	DPYS	FDXR	
ERCC3	ACO1	NT5C	
CFLAR	XPA	TK1	
UGT1A8	SLC31A1	TYMS	
UGT1A10	FPGS	RALBP1	
UGT1A9	AKR1C3	XRCC1	
UGT1A7	ERCC6	ERCC2	
UGT1A6	CYP2C8	ERCC1	
UGT1A5	ABCC2	PNKP	
UGT1A4	RRM1	TOP1	
UGT1A3	SMPD1	SOD1	
UGT1A1	GSTP1	CBR1	
DTYMK	SLCO1B1	CDC45L	
MLH1	TUBA1B	GSTT1	
MAP4	UNG	UPB1	
GPX1	TUBA3C	TYMP	
NR1I2	HMGB1	ATP7A	
UMPS	ATP7B	SMARCA1	
BCHE	ABCC4	CSAG2	
TUBB	MT2A	SLCO6A1	
ABCC5	TDP1	CES1	

Table S.3: FASTQ Statistics Summary. FASTQ statistics (mean±sd) for candidate gene and whole exome study samples are shown in this table. The library for candidate gene study was single end reads of variable lengths; whereas the exome study was paired end library of read length 75. Each sample was dedicated its own lane on an Illumina GAIIx – the average read counts per fastq are shown. It is not surprising that the paired end sequencing resulted in almost double the reads than single end sequencing. Exome study shows lower proportion of duplicates because of a more complex library as compared to the candidate gene study. %Ns in both the studies was negligible.
*Read Lengths varied with sequencing centers
**% Duplication is determined by looking at the first 2M reads in the fastq and counting how many times these reads are repeated in that fastq.

Study	**SE/PE**	**Read lengths***	**Reads per fastq**	**%duplicates****	**%N**
Candidate	SE	22, 31, 35, 36, 40	14307993 ± 7051967	17.24 ± 9.29	0.033 ± 0.034
Exome	PE	75	30793333 ± 3557653	0.82 ± 0.61	0.061 ± 0.069

Table S.4: Additional results of significant hits from association. Additional results of significant hits from association at FDR corrected threshold of 0.10 are shown in this table. All hits are known dbSNP variants. The allele frequency is reported from the low coverage pilot study of 1000 Genomes Project in CEU samples.
*Significant association with drug oxaliplatin at FDR threshold of 0.05

Drug(s)	SNP ID	SNP Position	Gene	CEU Frequency
Arsenic Trioxide	rs1846692 (C/T)	chr16:56671696	MT1A (near-gene 5)	C:0.783/T:0.217
Oxaliplatin	rs9922957 (C/G)	chr16:56672380	MT1A (near-gene 5)	C:0.883/G:0.117
	rs9922409 (A/G)	chr16:56672400	MT1A (near-gene 5)	C:0.067/T:0.933
	rs3749440 (C/T)	chr3:183702089	ABCC5 (UTR3)	C:0.417/T:0.583
Cytarabine, Oxaliplatin	rs2853742 (C/T)	chr18:657474	TYMS (near gene 5)	A:0.200/G:0.800
Bleomycin, Epirubicin	rs3749445 (A/G)*	chr3:183638506	ABCC5 (UTR3)	A:0.508/G:0.492
Epirubicin	rs562 (C/T)*	chr3:183637845	ABCC5 (UTR3)	C:0.517/T:0.483

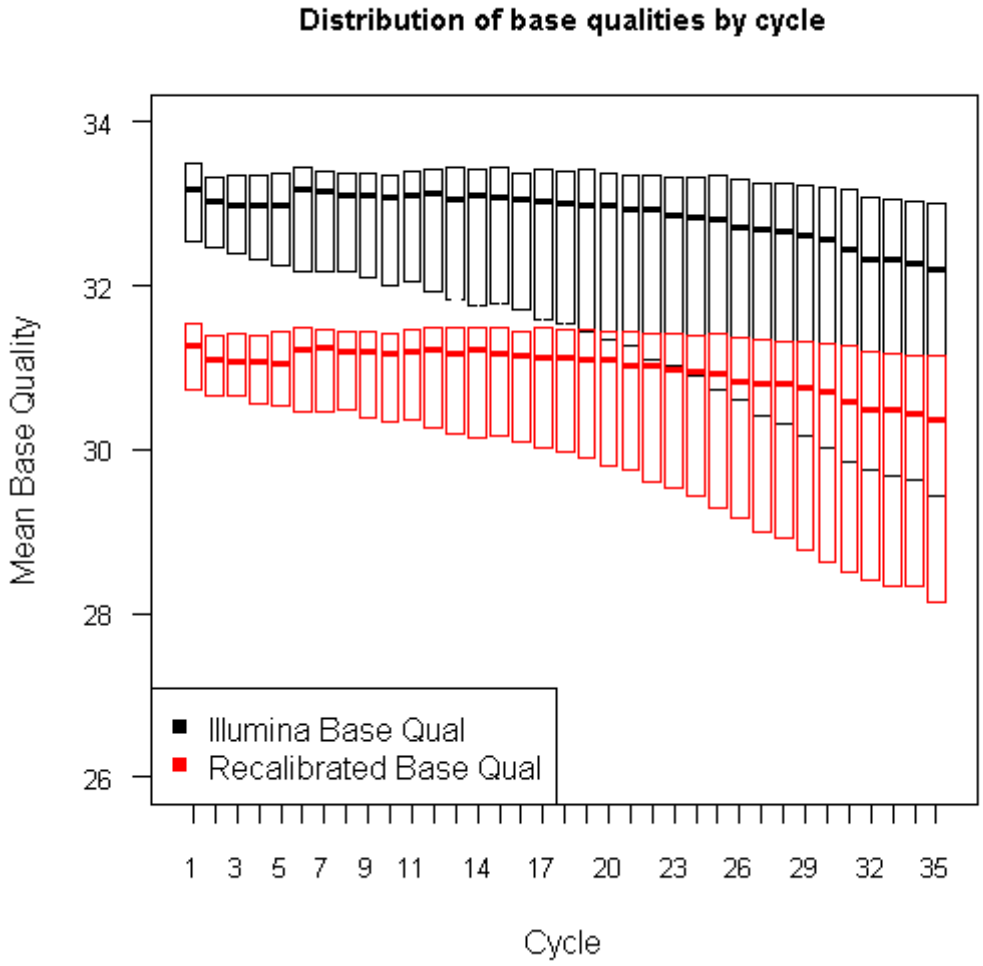

Figure S.1: Distribution of base qualities by cycle. In the top panel, Mean Base Quality scores at every sequencing cycle per FASTQ are plotted. Mean quality scores ranged anywhere between 27 and 34 before score calibration was done. The variance in the mean qualities of first few cycles is indicative of the variance introduced at different sequencing centers. The variance in mean qualities increases with sequencing cycles which is not unusual for Illumina short read sequencing. The bottom panel shows the recalibrated quality scores.

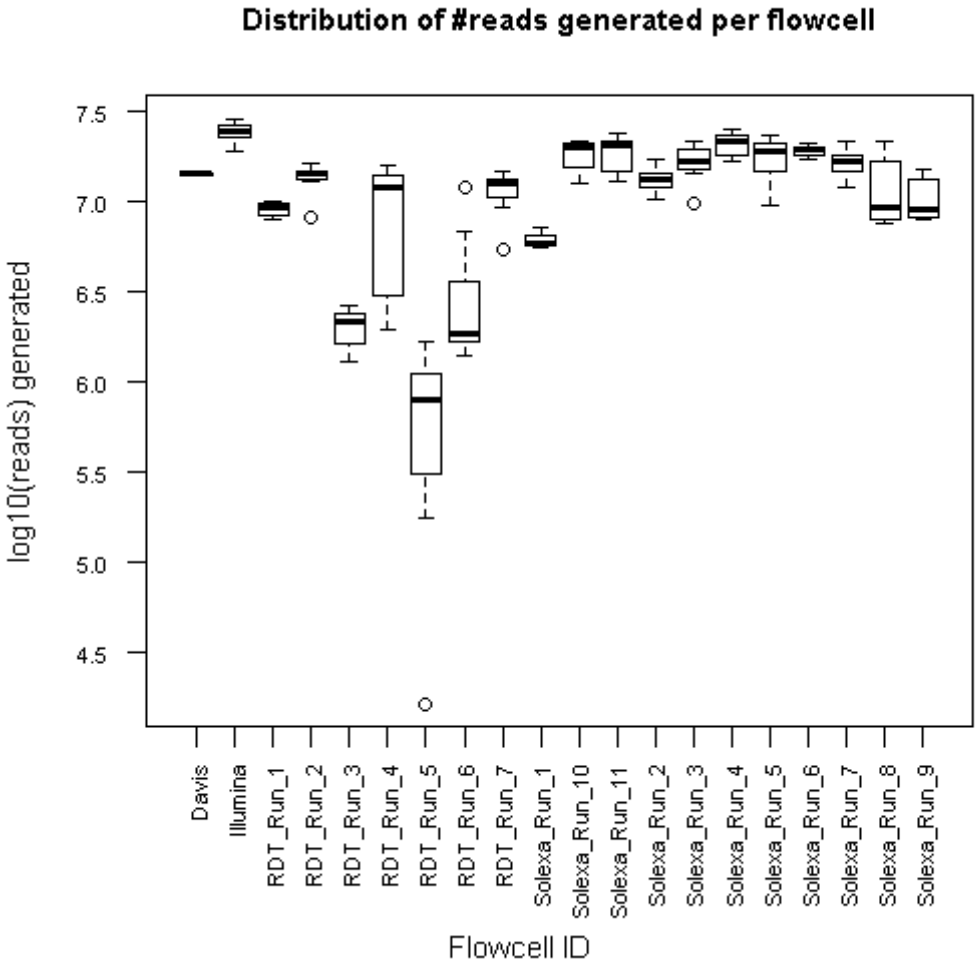

Figure S.2: Distribution of number reads generated at different centers. Each box plot represents the log scaled distribution of reads from different lanes of a particular flowcell. Read numbers varied by flowcell as well as centers.

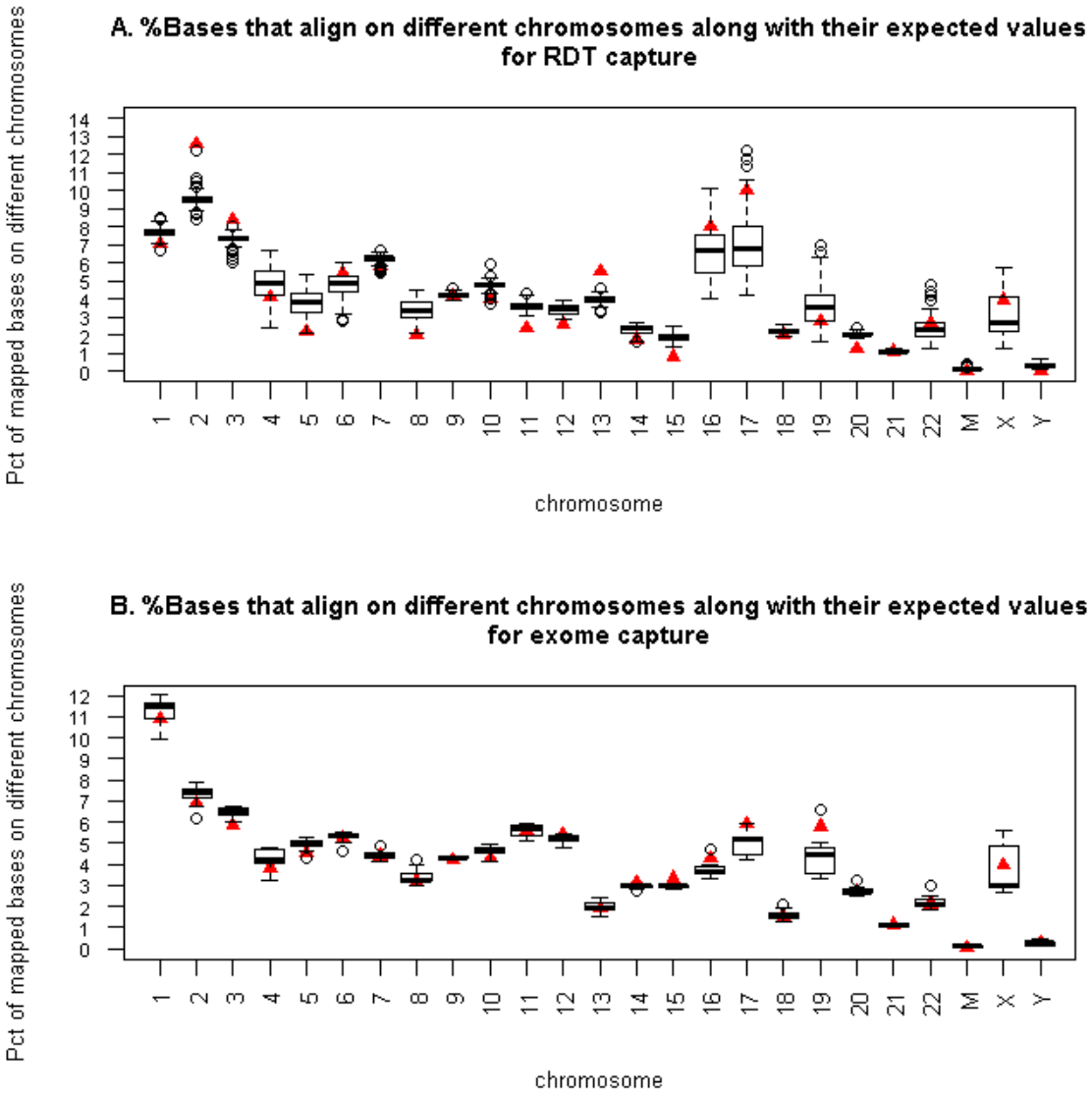

Figure S.3: Percent mapped bases on different chromosomes. Percentage of mapped bases along different chromosomes is plotted per sample. The solid triangles are the expected percent of mapped bases as calculated from the primer pairs designed for capture of candidate regions. Panel "A" shows the read distribution along different chromosomes for PCR based capture (RDT) for 134 samples and Panel "B" shows that for solution based capture (Roche NimbleGen) for 9 samples. The variability in RDT capture can be accounted partly to differences in FASTQ read counts across diverse sequencing centers.

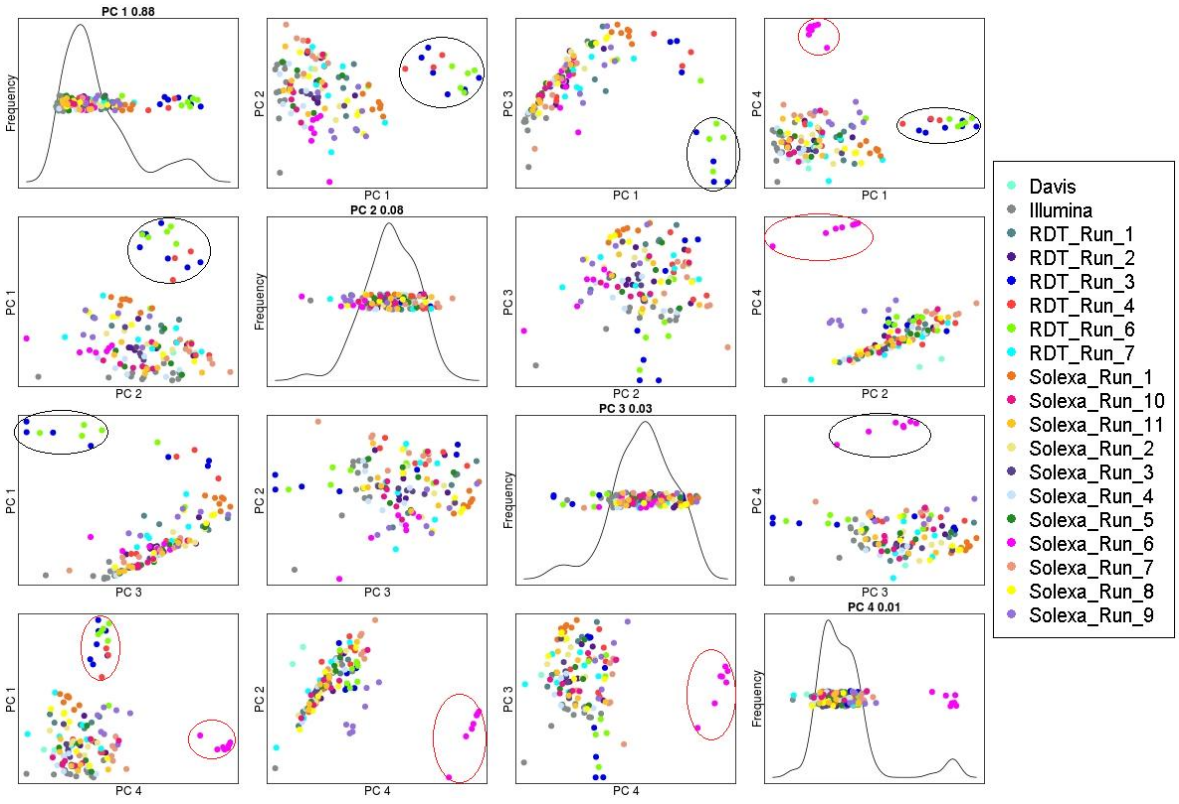

Figure S.4: Principal Component Analysis on summary statistics. Principal Component Analysis (that reduces dimensionality into primary axes of maximal variation) was done using various summary statistics. PCA was given a number of attributes like total number of reads, mean depth, coverage, uniformity of coverage, percent reads on target, percent bases on target, and percent amplicons captured. Using minimum number of these metrics, PCA tries to account for maximum variability in the original data. We can see that PCA shows a clear separation of capture qualities by flowcells. The outliers in blue, red, and green consist of samples from RDT Runs 3, 4 and 6 respectively; the magenta points are all samples from Solexa Run 6. Solexa Run 6 stands out because of low percent bases on target; RDT Runs show separation due to low read counts and low AURL scores.

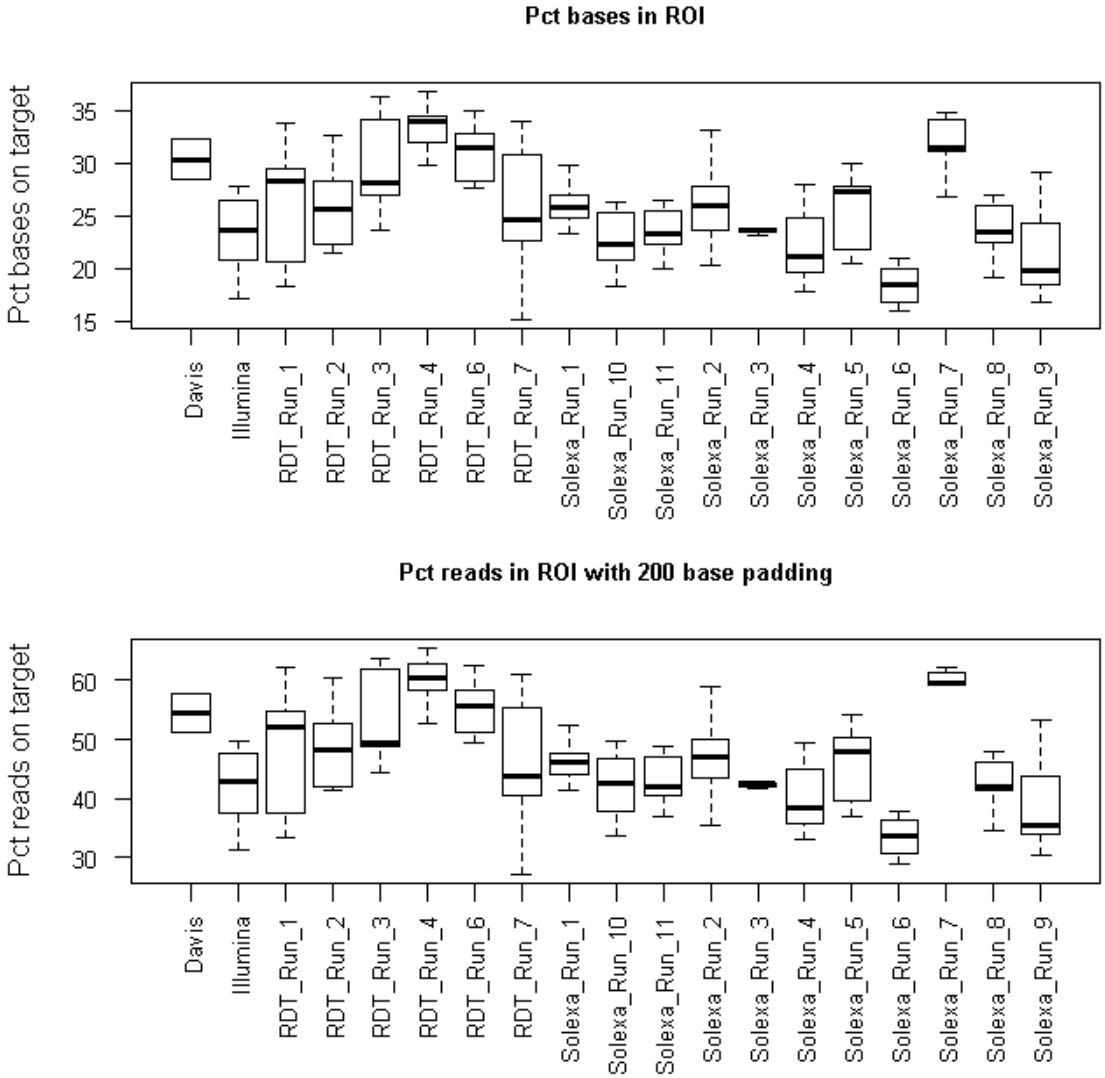

Figure S.5: Percent alignment on target. The plot in top panel shows the percentage of bases on target for all samples across different runs. The plot in bottom panel shows the percentage of reads that fall in the regions of interest (ROI) ± 200 bases.

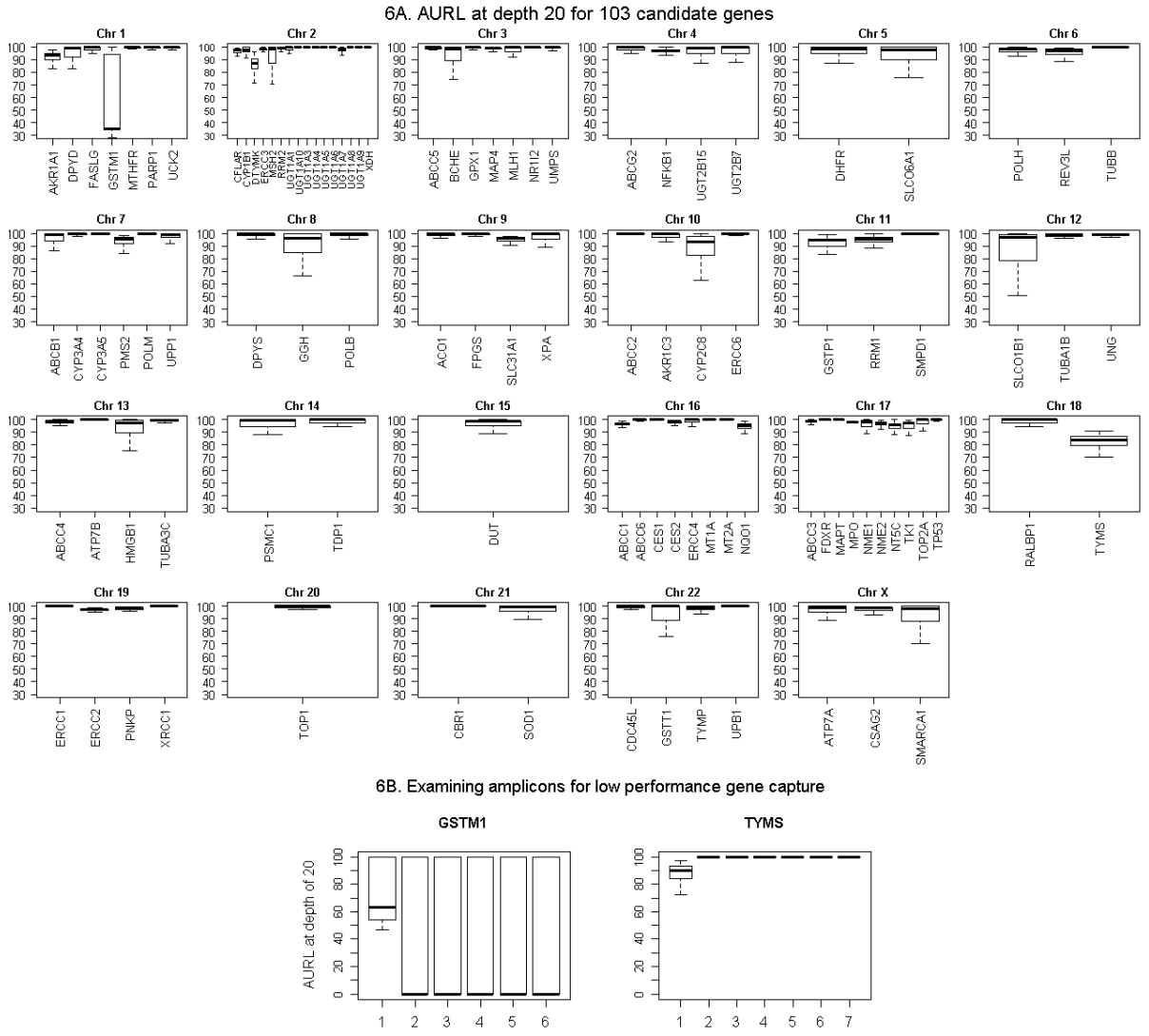

Figure S. 6: Target capture across genes for all samples. The plot in 6A summarizes all candidate genes intended for capture across various chromosomes along with their capture efficiency (AURL at depth 20). Gene GSTM1 on chromosome 1 clearly failed to be captured across many samples. Gene TYMS on chromosome 18 also shows low AURL values indicating primer failures for some exons. Plot 6B shows that all 6 amplicons for GSTM1 and 1 out of 7 for TYMS showed low efficacy across samples.

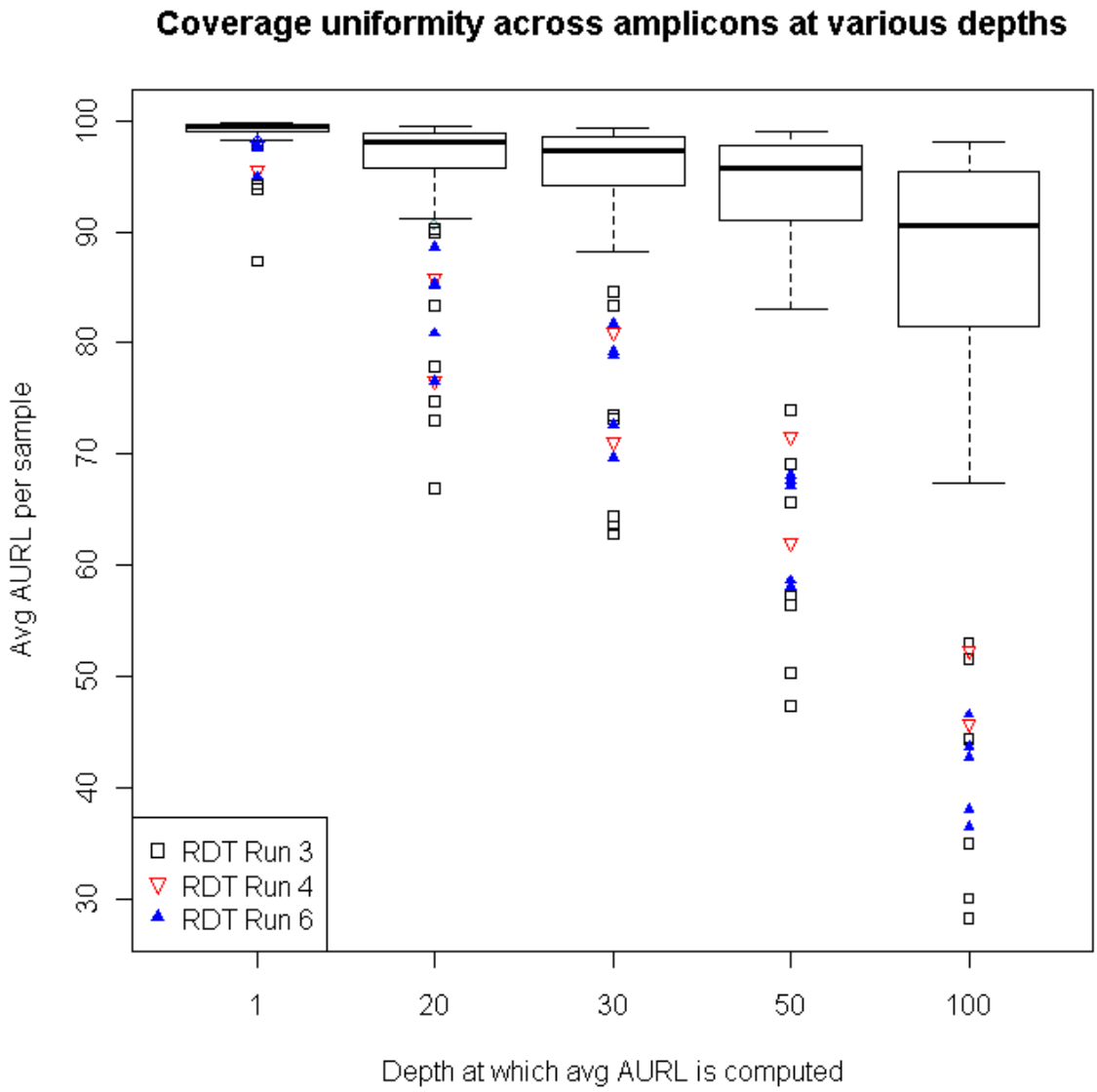

Figure S.7: Uniformity of coverage by sample. Overall uniformity of coverage at depths 1, 20, 30, 50, and 100 is plotted for every sample. The AURL is a function of depth and proportion of bases covered at that depth – the average AURL across all amplicons for a sample in plotted on y-axis. Seventy-five percent of samples had an average AURL greater than 95.77% at depth 20, indicating uniform coverage of most positions in regions of interest at this depth. Low AURL at depth 1 is a quick way to assess that a few amplicons failed to amplify in some samples. All the outlier points belong to flowcells from RDT Runs 3, 4 and 6.

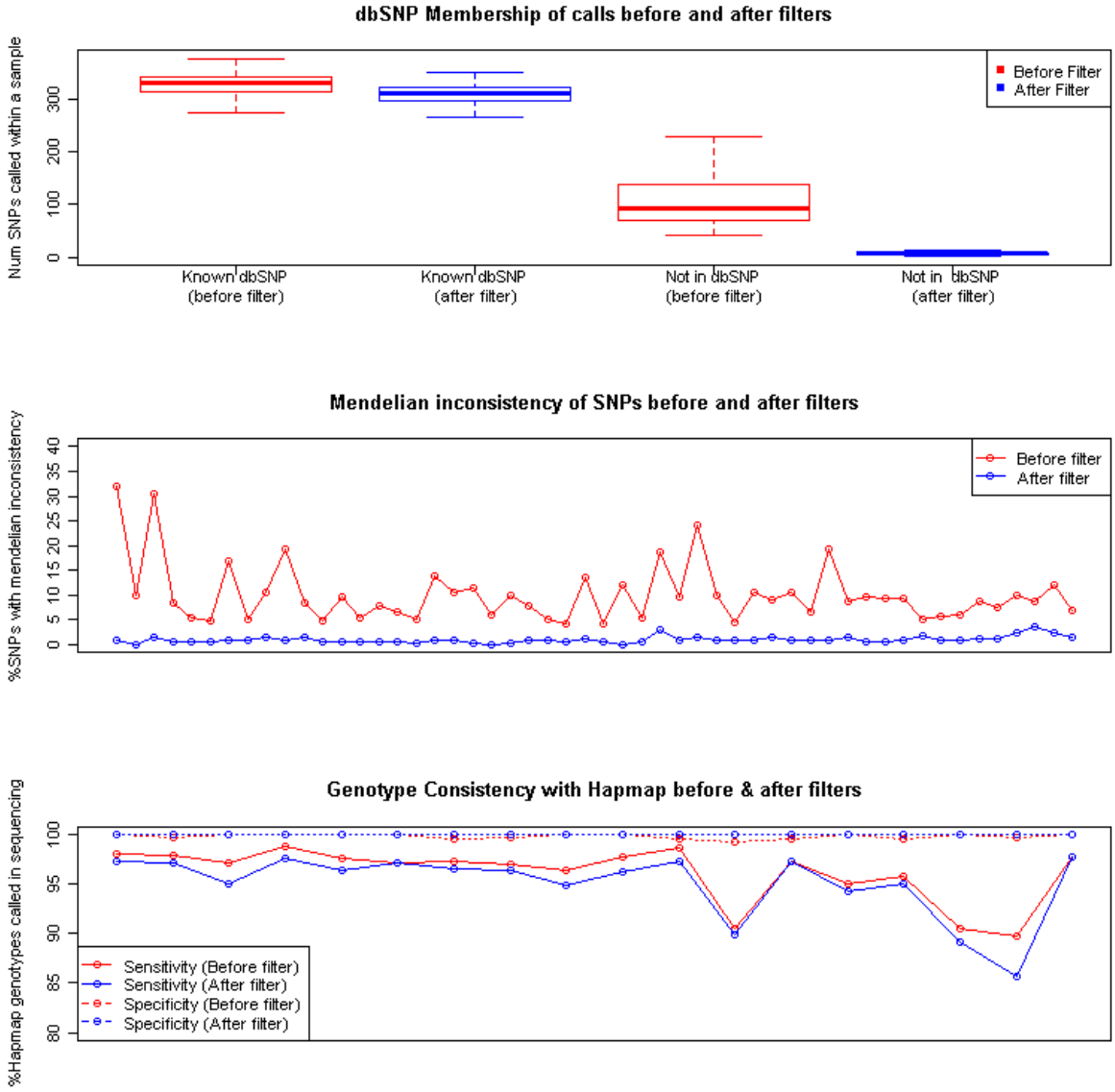

Figure S.8: Quality statistics for filtered SNPs. Quality of SNPs before and after filtering in terms of dbSNP membership, Mendelian consistency and Hapmap genotype consistency is shown. The top panel shows the dbSNP membership of SNPs that were retained – a significant decrease in non-dbSNP variants is seen. The middle panel plots percentage of SNPs showing Mendelian inconsistency in every trio – the inconsistency within all trios was below 4% after filtering. The bottom panel shows genotype consistency with Hapmap data available for 18 samples. Sensitivity is defined as the percent of variant calls made in Hapmap data, which were also made in sequencing data. Specificity is defined as the percent of homozygous reference calls in Hapmap data that were called non-variant in sequencing data.

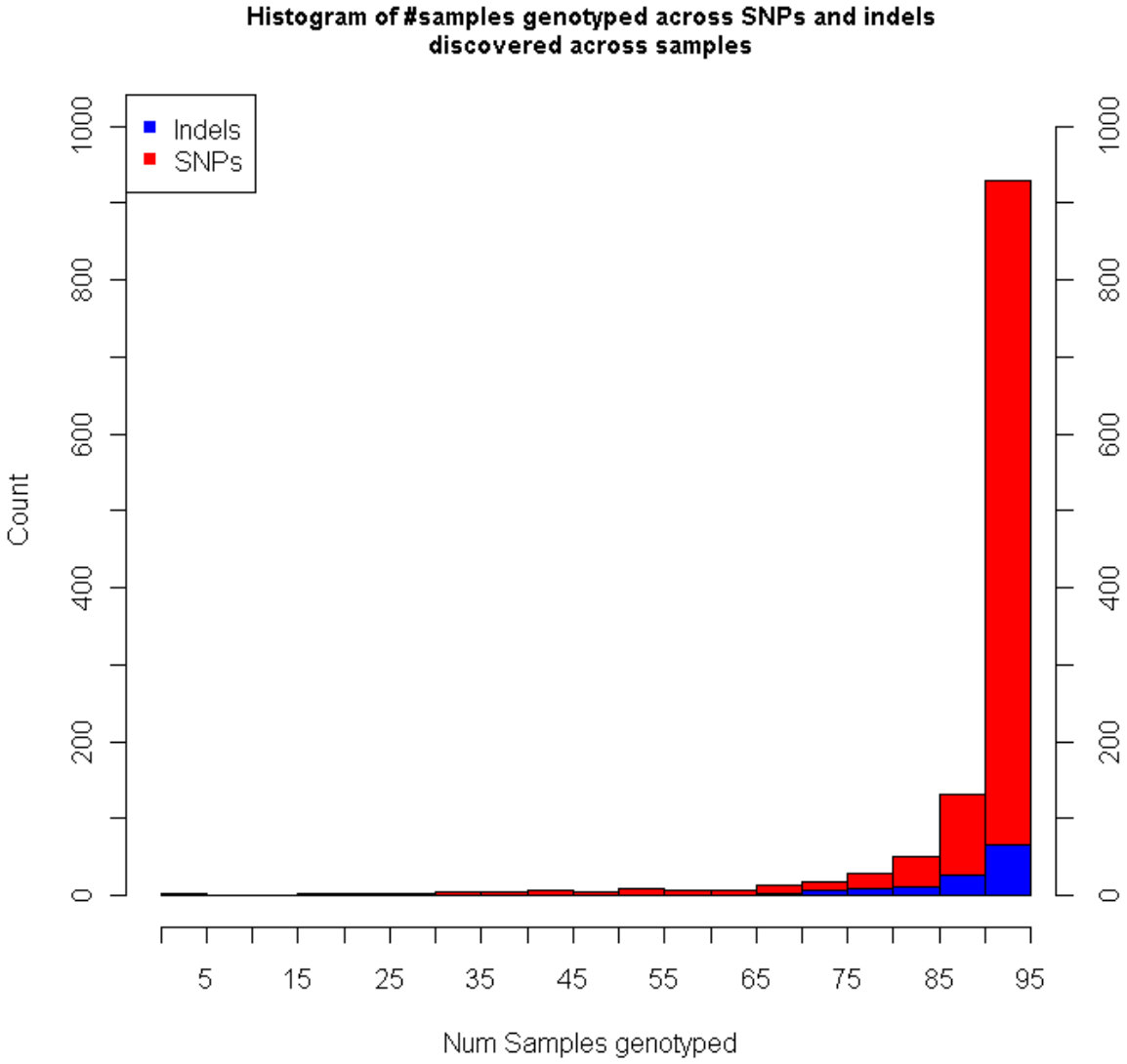

Figure S.9: Histogram of samples genotyped for a SNP. The red histogram depicts how many SNP markers were genotyped across the 94 cell lines – more than 90 samples were genotyped for approximately 960 markers. The blue histogram represents indel counts – 90 samples or more were genotyped at ~55% of indel marker positions observed in this study.

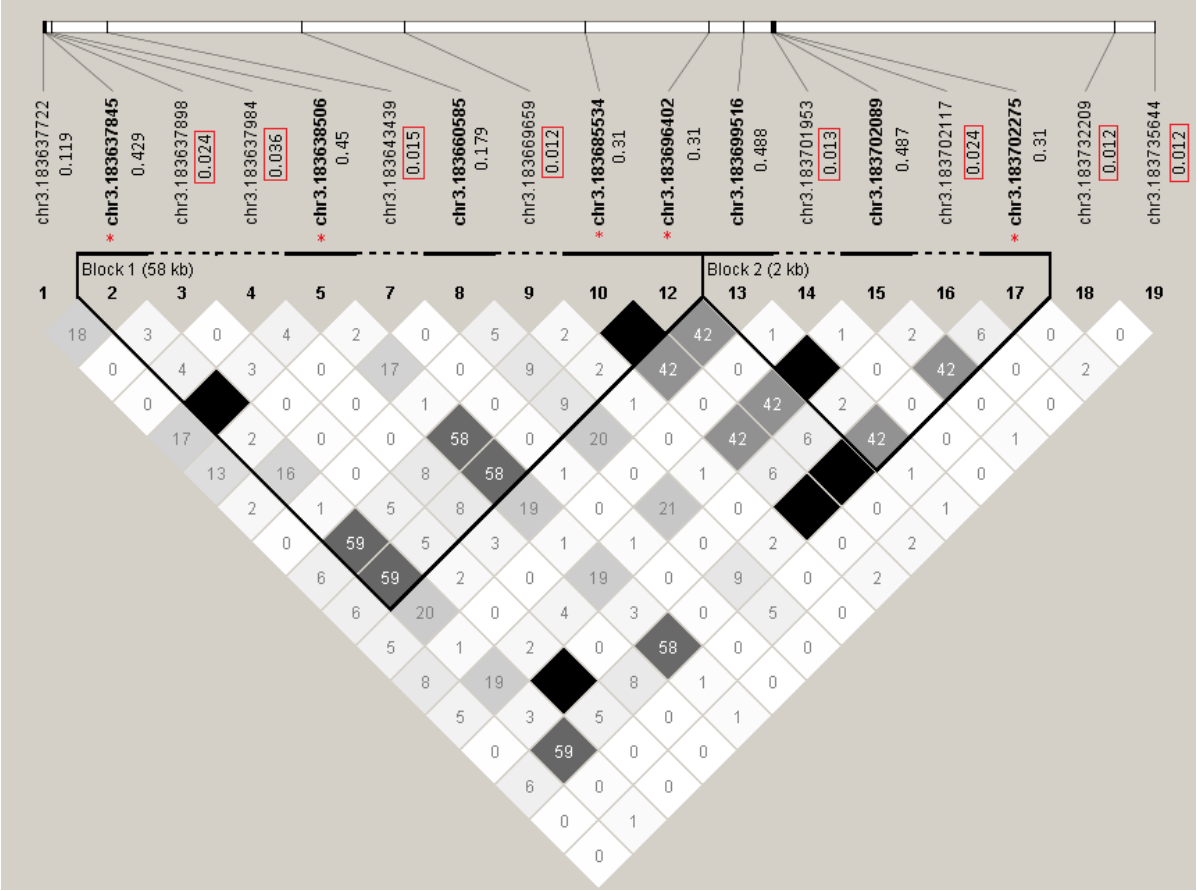

Figure S.10: Linkage Disequilibrium analysis for significant hits with Oxaliplatin. Haplotype block analysis with SNPs on chromosome 3 that are significant hits for the drug Oxaliplatin. Each marker is represented with the minor allele frequency as estimated in Haploview from the samples used to compute LD structure. The markers with asterisk are significant at FDR threshold of 0.05. Two haplotype blocks are defined interspersed with low frequency alleles (in red boxes). The numbers in the haplotype block are the r-squared statistic computed between two SNPs.

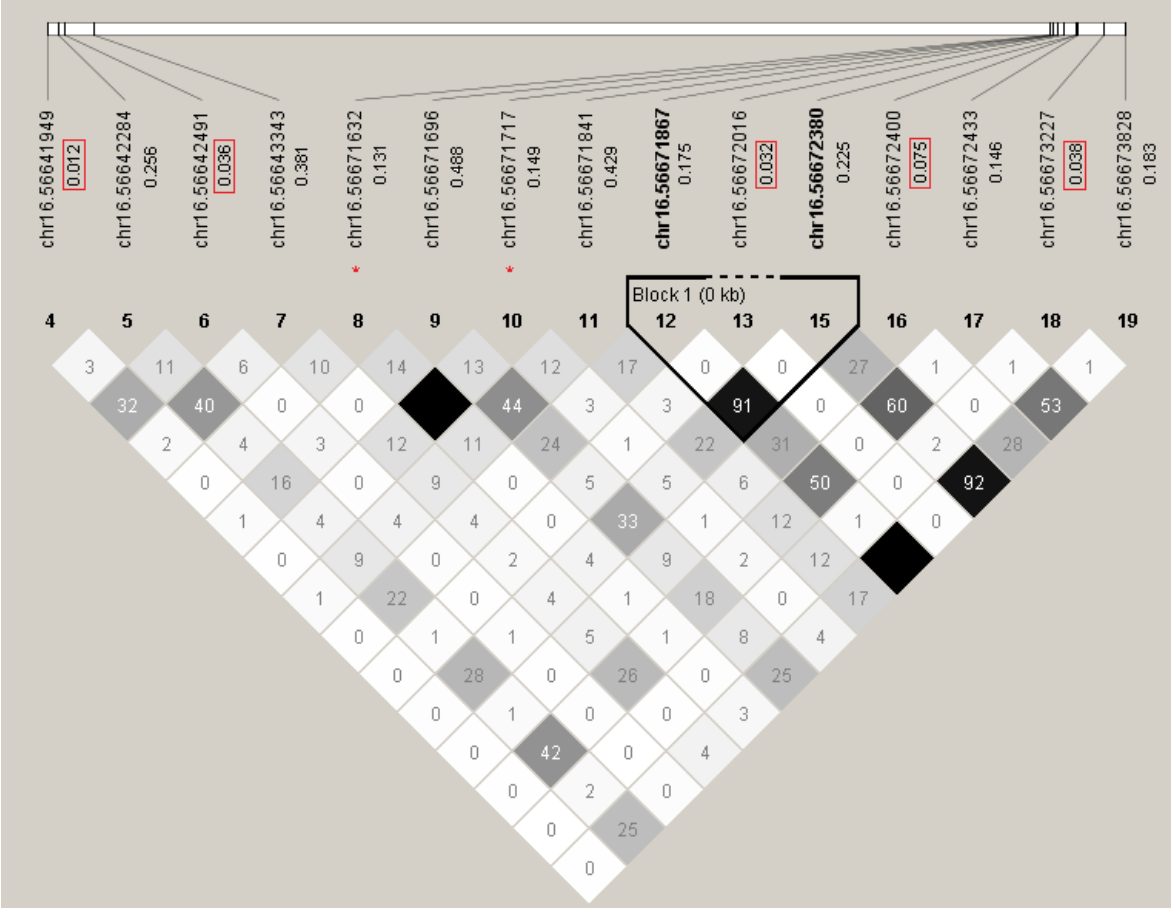

Figure S.11: Linkage Disequilibrium analysis for significant hits with Bleomycin. Haplotype block analysis with SNPs on chromosome 16 that are significant hits for the drug Bleomycin. The two significant hits show a perfect correlation with each other.