Title
Rare variant analysis for common diseases

Permalink
https://escholarship.org/uc/item/6081m00q

Author
Bhatia, Gaurav

Publication Date
2009

Peer reviewed|Thesis/dissertation
UNIVERSITY OF CALIFORNIA, SAN DIEGO

Rare Variant Analysis for Common Diseases

A Thesis submitted in partial satisfaction of the requirements for the degree
Master of Science

in

Computer Science

by

Gaurav Bhatia

Committee in charge:

Professor Vineet Bafna, Chair
Professor Yannis Papakonstantinou
Professor Pavel Pevzner

2009
The Thesis of Gaurav Bhatia is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

__________________________________________

__________________________________________

__________________________________________

Chair

University of California, San Diego

2009
# TABLE OF CONTENTS

Signature Page .......................................................... iii

Table of Contents ...................................................... iv

List of Figures ........................................................ v

Abstract of the Thesis ................................................. vi

Chapter 1  Introduction ................................................... 1

Chapter 2  Methods ...................................................... 5
  2.1  Modeling ......................................................... 5
  2.2  The RARECOVER method .................................... 6
  2.3  Parameters for rare variant simulation .................... 7
  2.4  Simulating populations ...................................... 8
    2.4.1  Simulating genotypes ............................... 10
  2.5  Reimplementing Proposed Methods ......................... 10
  2.6  CRESCENDO data ............................................. 10

Chapter 3  Results .................................................... 12
  3.1  Simulations .................................................... 12
  3.2  Obesity data results ....................................... 14
  3.3  Causal RVs and Metabolite expression levels ............. 16

Chapter 4  Discussion .................................................. 18

Bibliography .......................................................... 20
LIST OF FIGURES

Figure 2.1: The RARECOVER method for detecting locus association. In each step, the RV that maximizes the test statistic is chosen. The method stops when no further improvement is possible. $Q$ is a user defined parameter that can be adjusted to fine-tune the algorithm. While not explicit here, note that the computation of $X_{corr}$ requires the set of case-control genotypes. . . . . . . . . . . . . . 6

Figure 2.2: Generating Case-Control genotypes for RV simulation. Note that there is no explicit control of the number of causal RVs, but the choice of parameters helps to bound the number. . . . . . . . . . . . . . 9

Figure 3.1: The power of different methods for detecting rare variants. The $p$-value of association, based on permutation tests is presented for different values of $\rho \in [0.5, 0.8]$ and $R \in [0.1, 0.2]$. The power increases with an increase in $\rho, R$, or population size. . . . . . . . . . . . . . . . . . . 13

Figure 3.2: Comparisons between causal RVs, and RVs recovered by RARECOVER. The plot describes the number of causal RVs, RVs recovered, the intersection, and the fraction recovered. Note that a significant fraction of the causal RVs are recovered, with the ratio improving with larger populations. . . . . . . . . . . . . . . . . 14

Figure 3.3: FAAH locus association. The $x$-coordinate of each point corresponds to a single 5Kbp locus starting at that location. The $y$-coordinate corresponds a permutation based $p$-value. The most significant locus is $\sim$ 1Kbp upstream of the FAAH transcription start site. The region is part of an LTR element, which are known to carry regulatory signals. . . . . . . . . . . . . . . . . 15

Figure 3.4: MGLL Locus association. the most significant locus lies approximately 6Kbp upstream of the MGLL gene . . . . . . . . . 17

Figure 3.5: FAAH genotypes and AEA (anadamide) expression levels. The mean expression levels of AEA is significantly higher in cases containing RVs selected by RARECOVER, as compared to controls. 17
ABSTRACT OF THE THESIS

Rare Variant Analysis for Common Diseases

by

Gaurav Bhatia

Master of Science in Computer Science

University of California, San Diego, 2009

Professor Vineet Bafna, Chair

Genome-wide association (GWA) studies, which search for association between single, common genetic markers and a disease phenotype, have shown varying degrees of success. Recent studies suggest that multiple rare variants at different loci may act in concert to influence etiology of common diseases. However, current GWA studies depend upon the Common Disease Common Variant hypothesis and are not powered to account for disease-causing rare variants (CDRV).

We consider a simple CDRV model where a subset of rare variants at a locus is independently deleterious. These rare variants have modest penetrance (their presence increases disease likelihood significantly), but each explains only a small fraction of the diseased individuals. For this model, we describe an algorithm that efficiently computes a subset of rare-variants that best associate with the disease. We test our approach using extensive simulations. Neutral rare-variants are simulated using coalescent models, and causal variants are simulated under selection using Wright’s equation. Our method is robust in detecting associations with modest odds ratios in relatively small cohort studies. It significantly outperforms previously proposed strategies, and is robust to other CDRV models.
Our simulations suggest that associated regions with an odds ratio of 1.2 can be identified with a cohort study of 10000 individuals. Those with an odds ratio of 1.5 can be identified with a cohort study of 5000 individuals. Regions with stronger association (odds ratio > 2.2) can be identified with cohort studies of 1500 individuals or less.

The method was applied to a cohort of 289 individuals, 143 with high BMI and 146 with relatively low BMI. Two genes (187 kb) were sequenced in this cohort using next-generation sequencing. We identified two subsets of co-located variants that showed significant association with BMI (out of a total of 1088 rare variants). These subsets associated showed association with un-adjusted chi-square P-values of $3.6 \times 10^{-4}$ and $1.4 \times 10^{-4}$, and permutation based P-values of 0.004 and $8.04 \times 10^{-4}$, respectively. In comparison, the most significant association using single marker tests had a chi-square P-value of 0.002. Our results suggest that whole genome association using an analysis of rare variants is feasible.
Chapter 1

Introduction

The Common Disease, Common Variant (CDCV) hypothesis [Lan96, PC02, RL01] postulates that the etiology of common diseases is mediated by commonly occurring variants in a population. This has served as the basis for genome wide association (GWA) studies that test for association between single markers and the disease phenotype. Using genome-wide panels of common SNPs, GWA studies have been successful in identifying hundreds of statistically significant associations for many common diseases as well as several physiological traits [Con07, FTW+07, HTM+07, MPK+07]. Nevertheless, the success of GWA studies has been mixed. Significant genetic loci have not been detected for several common diseases that are known to have a strong genetic component [Con07]. Additionally, for many common diseases, associations discovered in GWA studies can account for only a small fraction of the heritability of the disease. While many factors could potentially confound GWA studies, we focus on the possibility that multiple, rare variants may act in concert to influence disease etiology.

This Common Disease, Rare Variant (CDRV) hypothesis has been the topic of much recent debate [SMFT09], but has shown promise in explaining disease etiology in multiple studies. Rare variants (RVs) have been implicated in reduced sterol absorption and, consequently, lower plasma levels of LDL [CPF+06, CKP+04]. Additionally, RVs have been shown to increase risk of colorectal cancer [FWW+04]. While some studies have shown RVs to increase risk, others have indicated that RVs also act ‘protectively’, reducing risk of common disease. Multiple RVs in renal
salt handling gene have shown association with reduced renal salt resorption and reduced risk of hypertension\cite{JFO+08}. Additionally, rare mutations in IFIH1 have been shown to act protectively against type 1 diabetes\cite{NWR+09}.

The studies above, and others were focused on re-sequencing of coding regions of candidate genes using Sanger sequencing (see Table 1 in Schork et al.\cite{SMFT09} for a summary). Recent technological advances in DNA sequencing have made it possible to resequence large stretches of a genome in a cost-effective manner. This is enabling large-scale studies of the impact of RVs on complex diseases. However, several properties of rare variants make their genetic effects difficult to detect with current approaches. Bodmer and Bonilla provide an excellent review of the properties of RVs, and the differences between rare, and common variant analyses\cite{BB08}. As an example, if a causal variant is rare (MAF 0.001-0.05), and the disease is common, then the allele’s Population-Attributable-Risk (PAR), and consequently the odds-ratio (OR), will be low. Additionally, even highly penetrant RVs are unlikely to be in Linkage Disequilibrium with the incidence vector for a common disease. Therefore, single-marker tests of association, which measure LD based correlation, are likely to have low power. If the CDRV hypothesis holds, a combination of multiple RVs must contribute to population risk. In this case, there is a challenge of detecting multi-allelic association between a locus and the disease.

Methods to detect such associations are only just being developed. A natural approach is a collapsing strategy, where multiple RVs at a locus are collapsed into a single variant. Such strategies have low power as ‘causal’ and neutral RVs are combined (See for example, Li and Leal\cite{LL08}). Madsen and Browning have recently proposed a weighted sum statistic to detect loci in which disease individuals are enriched for rare variants\cite{MB09}. In their approach, variants are weighted according to their frequency in the unaffected sample, with low frequency variants being weighted more heavily. Each individual is scored as a sum of the weights of the mutations carried. The test then determines if the diseased individuals are weighted more heavily than expected in a null-model. Madsen and Browning show that with 50\% of variants in a group being causal and a combined odds ratio > 15,
the weighted sum statistic detects associations with high power. While effective, this approach depends upon a high proportion of causal rare variants and strong penetrance to detect significant association.

The Combined Multivariate and Collapsing Method (CMC), proposed by Li and Leal, combines variants into groups based upon predefined criteria (e.g. allele frequency, function)[LL08]. An individual has a ‘1’ for a group if any variant in the group is carried and a ‘0’ otherwise. The CMC approach then considers each of the groups in a multivariate analysis to explain disease risk. This combination of the collapsing approach and multivariate analysis results in an increase of power over single-marker and multiple marker approaches. However, as Li and Leal point out, the method relies on correct grouping of variants. The power is reduced as functional variants are excluded and non-functional variants are included in a group. Assignment of SNPs to incorrect groups may, in fact, decrease power below that attainable through single marker analysis.

Here, we focus on a model-free method, RareCover, that collapses only a subset of the variants at a locus. Informally, consider a locus \( L \) encoding a set \( S \) of rare variants. RareCover associates \( L \) with a phenotype by measuring the strongest possible association formed by collapsing any subset \( S' \subseteq S \) of variants at \( L \). At first glance, such an approach has many problems. First, selecting an optimal subset of SNPs is computationally intensive, scaling as \( 2^{|S|} \). We show that a greedy approach to selecting the optimal subset scales linearly, making it feasible to conduct associations on a large set of candidate loci.

A second confounding factor is that the large number of different tests at a locus increase the likelihood of false association. However, extensive simulations show otherwise. Our results suggest that moderately penetrant alleles \( (RR \geq 1.25) \) with small PAR (\( \leq 20\% \)), moderately sized cohorts (\( \sim 500 \) cases and \( \sim 500 \) controls) are sufficient to detect significant associations with the disease. For the purposes of comparison, RareCover can detect associations with an odds-ratio \( > 1.5 \) with moderately small cohort studies. This compares well with the current power of single-marker GWA studies.

We also applied RareCover to the analysis of two genes, FAAH, and
MGLL, in the endocannabinoid pathway, which is an important mediator of a variety of neurological functions\cite{DMBDP04, DM09}. Endocannabinoids acting upon CB1 receptors in the brain, the gastrointestinal tract, and a variety of other tissue have been shown to influence food intake and weight gain in animal models of obesity. Using a selective endocannabinoid receptor (CB1) antagonist, SR141716 (rimonabant; Sanofi-Synthelabo) leads to reduced food intake in mice. Correspondingly, elevation of leptin levels have been shown to decrease concentrations of endogenous CB1 agonists, Anandamide, and 2-AG in mice, thereby reducing food intake\cite{DMGW01}. The FAAH and MGLL enzymes serve as regulators of endocannabinoid signalling in the brain\cite{CGM96}, by catalyzing the hydrolysis of endocannabinoids including anandamide (AEA), and 2-AG. Gene expression studies in lean and obese women show significantly decreased levels of AEA and 2-AG, as well as overexpression of CB1 and FAAH in lean, as opposed to obese women\cite{EBF05}. While evidence points to a genetic association of these loci with obesity, multiple recent studies using common SNPs in the FAAH region have failed to confirm an association with obesity\cite{JAA07, EDLD08, MRB08, LMF09}. A Pro129Thr polymorphism was tentatively associated with obesity in a cohort of Europe and Asian ancestry, but has not been replicated in other data\cite{SWGB05}.

We ask if multiple, rare alleles at these loci associated with obesity. We have used unpublished data from Frazer and colleagues, where the FAAH (31Kbp) and MGLL (156KBp) regions were re-sequenced using next generation technologies. The resequencing was done with individuals from the extremes of BMI distribution (the CRESCENDO cohort NCT00263042), and identified a number of common, and rare variants in the region. We applied RARECOVER to determine if allelic heterogeneity mediated the genetic effects of FAAH, and MGLL. RARECOVER identified a single region at each locus with permutation adjusted p-values of $0.0036$ and $6.50 \cdot 10^{-4}$. In each case, the significant locus was immediately upstream of the gene, consistent with a regulatory function for the rare variants.
Chapter 2

Methods

2.1 Modeling

We define a locus as a genomic region of fixed nucleotide length. Let $S$ denote the set of RVs in the locus. We abuse notation slightly by using $S$ to also denote the locus itself. A case-control study at $S$ includes a set of individual genotypes. For genotype $I$, and RV $s \in S$, let $I_s \in \{0, 1, 2\}$ denote the number of minor alleles that genotype $I$ carries for variant $s$. We extend the notation to a subset $C \subseteq S$ of SNPs, as $I_C = \sum_{s \in C} I_s$. For a subset $C \subseteq S$, denote a locus-variant $A_C$ as follows: individual $I$ has the allele $A_C = 1$ if and only if $I_C > 0$. Otherwise, $A_C = 0$. Let $D = 1$ (respectively, $D = 0$) represent the case (respectively, control) status of an individual.

For an individual chosen at random, and $C \subseteq S$, let $X_{\text{CORR}}(A_C, D)$ denote the association test statistic between the locus-variant $A_C$ and the disease status $D$ (Ex: Pearson’s $\chi^2$). Using this notation, the collapsing strategy described by Li and Leal[LL08] uses the test-statistic $X_{\text{CORR}}(A_S, D)$ to associate a locus $S$ with the disease. Instead, we define the association statistic for locus $S$ by

$$X_{\text{CORR}}(S, D) = \max_{C \subseteq S} X_{\text{CORR}}(A_C, D) \quad (2.1)$$

The objective of rare-variant analysis is to identify loci $S$ for which $X_{\text{CORR}}(S, D)$ is significant.
2.2 The RareCover method

The naive computation for $X_{corr}(S, D)$ needs $2^{|S|}$ computations. A reduction from the MAX-COVER problem can be used to show that the problem is NP-hard, indicating that no provably efficient algorithm is likely [GJ79]. Similar reductions can also be used to show the hardness result for a variety of other proposed association statistics. Therefore, we use a heuristic that select a set $C$ with a high $X_{corr}$ score. RareCover is a greedy approach. In each step, (see Figure 2.1), we select the variant that adds the most to the statistic, until no further improvement is possible.

**procedure** RareCover($S, Q$)

Set $C = \emptyset$, $u = \emptyset$

repeat

Set $C = C + \{u\}$

Select $u \in S - C$ that maximizes $X_{corr}(S_{C+\{u\}}, D)$

while \((X_{corr}(S_{C+\{u\}}, D) - X_{corr}(S_C, D) \geq Q)\)

return $X_{corr}(S_C, D)$.

Figure 2.1: The RareCover method for detecting locus association. In each step, the RV that maximizes the test statistic is chosen. The method stops when no further improvement is possible. $Q$ is a user defined parameter that can be adjusted to fine-tune the algorithm. While not explicit here, note that the computation of $X_{corr}$ requires the set of case-control genotypes.

While RareCover should improve upon the collapsing statistic, $X_{corr}(A_S, D)$ it is not clear that it improves the power, as the statistic will also increase for non-associated loci. We compute significance by applying RareCover to permutations of the case-control genotypes. To save time, we run permutations in a data-driven fashion. As an example, let RareCover($S, D$) = $t$. To compute a p-value, we run a maximum of $k$ iterations, but stop as soon as we obtain 2 samples for which the RareCover statistic exceeds $t$.

$$p\text{-val}(t) = \begin{cases} \frac{2}{k'} & \text{if } 2 \text{ samples exceed } t \text{ in } k' < k \text{ iterations} \\ \frac{1}{k} & \text{otherwise} \end{cases}$$ (2.2)
The maximum number of iterations, \( k \), is determined based on desired level of significance.

### 2.3 Parameters for rare variant simulation

Consider a locus with a set \( S \) of rare-variants. Let a subset \( C \) of RVs be causal, in the sense that a mutation at any \( s \in C \) increases the likelihood of disease. For an individual, \( I \), we use \( A_C \) and \( \bar{A}_C \) as short-forms of the events \( A_C = 1 \) (or, \( I_C > 0 \)), and \( A_C = 0 \), respectively. Similarly, events \( D, \bar{D} \) reflect case-control status for the individual. We work with the following 3 parameters for power calculations:

1. **Disease prevalence** in the population, denoted by \( P_D \).
2. **Penetrance** of the locus, denoted by \( \rho = \Pr(D|A_C) \)
3. **Locus-PAR**, denoted by \( R = \Pr(A_C|D) \)

Note that the PAR for a variant is often described by the following (Ex: Bodmer and Bonilla, 1999[BB08])

\[
R = \frac{K - y}{K} = 1 - \frac{y}{K} \tag{2.3}
\]

where \( K \) is the number of individuals with the phenotype, and \( y \) is the number of individuals that show the phenotype, but do not have the variant allele. In our terminology

\[
R = 1 - \frac{\Pr(\bar{A}_C \cap D)}{P_D} = 1 - \Pr(\bar{A}_C|D) = \Pr(A_C|D) \tag{2.4}
\]

The choice of these parameters is intuitive as we expect an RV to have moderate penetrance, but very low PAR (\( \Pr(A_s|D) \)). However, the multiple RVs in \( C \) have roughly additive effect, leading to moderate locus-PARs. These parameters are tightly related to other, more common measures of locus association, such as the Odds-Ratio (OR), as shown below:

\[
\text{OR}(S) = \frac{\Pr(D|A_C)}{\Pr(D|\bar{A}_C)} / \frac{\Pr(D|\bar{A}_C)}{\Pr(D|A_C)} = \frac{\rho}{1 - \rho / \Pr(D|A_C)}
\]
To compute $\Pr(D|\bar{A}_C)$, we start with a Bayesian relation for computing the likelihood of a genotype containing a causal RV as

$$\Pr(A_C) = \frac{\Pr(A_C|D)\Pr(D)}{\Pr(D|A_C)} = \frac{RP_D}{\rho}$$  \hspace{1cm} (2.5)$$

Then,

$$\Pr(D|\bar{A}_C) = \frac{\Pr(\bar{A}_C|D)P_D}{\Pr(A_C)} = \frac{(1-R)P_D}{\Pr(A_C)} = \frac{(1-R)P_D}{1 - \left(\frac{RP_D}{\rho}\right)}$$ \hspace{1cm} (2.6)$$

and,

$$\Pr(\bar{D}|\bar{A}_C) = 1 - \Pr(D|\bar{A}_C).$$ \hspace{1cm} (2.7)$$

### 2.4 Simulating populations

We simulate multiple case-control studies over a range $P_D, \rho, R$. A simulation of $N$ individuals begins with the division of the individuals into $\left\lceil \frac{N}{2} \right\rceil$ cases and $\left\lfloor \frac{N}{2} \right\rfloor$ controls. Once this is done two additional steps take place.

1. Generate a set of RVs for the simulated locus containing causal, and neutral RVs.
2. Simulate the genotypes for each individual.

We start by generating causal RVs. As RVs do not show high LD, we can model the population by generating each RV independently. We adapt Pritchard’s argument that the frequency distribution of rare, deleterious, RVs must follow Wright’s model under purifying selection[Pri01]. Therefore, the allele frequencies $p$ are sampled according to:

$$f(p) \propto p^{(\beta_S-1)}(1-p)^{(-1)(\beta_N-1)}e^\sigma(1-p)$$ \hspace{1cm} (2.8)$$

where,

- $p$, allelic frequency
- $\sigma$, selection coefficient
- $\beta_S$, rate of mutation from normal allele to causal
- $\beta_N$, rate of repair from causal allele to normal
We choose $\sigma = 30.0, \beta_S = 0.2, \beta_N = 0.002$[Pri01]. Note that we do not control the number of causal RVs, $|C|$, directly, in our simulation. Recall that

$$\Pr(A_C|D) = 1 - \Pr(\bar{A}_C|D) = 1 - \prod_{s \in C} \Pr(\bar{A}_{\{s\}}|D)$$

Further,

$$\Pr(\bar{A}_{\{s\}}|D) = 1 - \Pr(A_{\{s\}}|D) = 1 - \left( \frac{\Pr(A_{\{s\}}) \Pr(D|A_{\{s\}})}{\Pr(D)} \right) = 1 - \frac{\pi_s \rho}{P_D}$$

Therefore, setting a value for $R$ limits the size of the causal RV.

$$R = \Pr(A_C|D) = 1 - \prod_{s \in C} \left( 1 - \frac{\pi_s \rho}{P_D} \right)$$

(2.9)

Figure 2.2 describes the method to generate causal RVs. To generate neutral RVs, we use Fu’s model of allele distributions, based on coalescent theory[Fu95]. This states that the number of mutations that affect $i$ individuals in a population with mutation rate $\theta$ is given by $\theta/i$. That is, the number of mutations that affect 2 haplotypes is half the number of mutations that affect only a single haplotype. For the purposes of our simulation we use $\theta = 10.0$.

**procedure**  
SIMULATERV()

Set $P = 1$
Set $C = \emptyset$
Repeat
Sample $\pi_s$ from Wright’s distribution
Generate a SNP $s$ with MAF $\pi_s$
$C = C + \{s\}$
$P = P * \left( 1 - \frac{\pi_s \rho}{P_D} \right)$
while $P > (1 - R)$

Figure 2.2: Generating Case-Control genotypes for RV simulation. Note that there is no explicit control of the number of causal RVs, but the choice of parameters helps to bound the number.
2.4.1 Simulating genotypes

For both cases and controls, each RV is sampled independently. For non-causal variants $s \in S - C$, the probability of picking a minor allele is $\pi_s$, for both case and control individuals. To sample alleles from causal SNPs, recall that under the union model, $\Pr(D|A_{\{s\}}) = \Pr(D|A_C) = \rho$ for all $s \in C$. Therefore, the minor allele frequencies are given by

$$\Pr(A_{\{s\}}|D) = \frac{\rho \pi_s}{P_D}$$
$$\Pr(A_{\{s\}}|\bar{D}) = \frac{(1 - \rho) \pi_s}{(1 - P_D)}$$

We assume HW equilibrium to sample genotypes for case and control individuals.

2.5 Reimplementing Proposed Methods

For the purposes of comparison we reimplemented statistics based upon the collapsing statistic proposed by Li and Leal[LL08] as well as the weighted-sum statistic used by Madsen and Browning[MB09]. Both publications discuss the separation of variants into groups based upon function (i.e. non-synonymous coding SNPs) or other property. However, because we are performing our studies on model free, unannotated data, we do not perform any such grouping.

We simulate our tests over a variety of parameters to better understand how RARECOVER performed in comparison to currently published methods under differing circumstances. Code for all methods is available upon request from the authors.

2.6 CRESCENDO data

In a recently submitted study, Frazer and colleagues resequenced a 31Kbp region from the FAAH gene and a 156Kbp region spanning the MGLL gene in 289 individuals of Caucasian ancestry from the CRESCENDO cohort (NCT002063042).
The individuals are selected from the extremes of Body mass index (BMI) distribution. Approximately half of the individuals had a Body mass index (BMI) below 30 while the remaining individuals had a BMI of 40 or greater.

Variants were detected, and distinguished from sequencing errors using proprietary methods. The resequencing identified 165 + 923 RVs (MAF ≤ 0.1) and also 63 + 262 common SNPs (MAF > 0.1). For our analysis, only the RVs were used. The data-set is available upon request from the authors.
Chapter 3

Results

3.1 Simulations

We simulated cases and controls (as described in Methods) for a collection of population sizes, ranging from 40 to over 5000 individuals with equal numbers of cases and controls. The MAF for rare variants ranged from $10^{-4}$ to $10^{-1}$. Throughout, we assume the disease prevalence in the population to be $P_D = 0.4$, and the PAR was set to either $R = 0.1$ or $R = 0.2$. The penetrance, $\rho$ was varied in the interval $[0.5, 0.8]$, corresponding to OR values of 1.56-7.25. While the disease prevalence is high, we note that $\frac{\rho}{P_D}$ is a lower bound on relative risk. Reducing $P_D$ would increase the relative risk, only making association easier.

While Genome wide significance is difficult to assess, we used a permutation based $p$-value of $10^{-4}$ as significant. This is based on the assumptions of 100 candidate genes for a phenotype, and about 100 loci per gene, based on a span of 10KBp, and an RV every 100bp. The results of the simulation are shown in Figure 3.1. For $\rho = 0.8$, $R = 0.1$, corresponding to an OR= 6.56, we obtained a $p$-value better than $10^{-4}$ for population of 1280 individuals. By contrast, the $p$-values obtained for the collapsing strategy, and MB, are $\sim 10^{-1}$, and $10^{-2}$, respectively. Even at modest penetrance values ($\rho = 0.5$), it is possible to detect RV based associations with a larger cohort. Not surprisingly, all methods improve with increasing values of penetrance, PAR, and population-size. Our results point to the power of rare-variant analysis at loci with modest PAR.
Figure 3.1: The power of different methods for detecting rare variants. The $p$-value of association, based on permutation tests is presented for different values of $\rho \in [0.5, 0.8]$ and $R \in [0.1, 0.2]$. The power increases with an increase in $\rho$, $R$, or population size.

An advantage of the RareCover approach is that it does not depend upon MAF, or the density of RVs in a region, as it seeks to combine the effects of multiple RVs that associate, while discarding the ones that do not associate. By contrast, other methods combine all RVs, albeit with different weights. While RareCover does not recover all of the causal RVs, it always recovered a significant subset of the causal RVs in our simulations. See Figure 3.2, which summarizes the results for $\rho = 0.8, R = 0.1$. $C, C^*$ correspond to the set of RVs detected by RareCover, and the simulated, causal RVs, respectively. Thus, $\frac{|C\cap C^*|}{|C^*|}$ corresponds to the fraction of causal RVs recovered. With larger population sizes, more than 50% of the RVs are recovered. A somewhat unexpected aspect is that the number of causal RVs, ($|C^*|$), (and also, $|C|$) increases with an increasing population size. For larger populations, we can recover a larger number of the low frequency variants, and the causal set has a larger mix of low frequency RVs. As we only consider RVs with MAF $> 10^{-4}$, the number saturates by 10K individuals.
3.2 Obesity data results

Single marker association tests identified two regions (both in the MGLL gene) showing moderate association with BMI. We speculated if multiple rare variants in the two genes could explain the difference in BMI between the two groups.

We applied RareCover to overlapping loci of length 5Kbp over the region. The permutation based $p$-value for each locus is plotted in Figure 3.3. The analysis suggests that a $\sim 5000$bp region is enriched with significantly associating loci. The locus with the most significant association, located $[-5705 \cdots -716]$ upstream of the FAAH TSS, contains 33 RVs. RareCover selected a subset of 16 RVs, with a union variant that appears in 23 cases, and 0 controls (permutation $p$-value 0.00364). Analyzing the locus for functional significance, the locus falls within a retroviral Long-Terminal-Repeat (LTR). Insertion of retroviral elements, followed by adaptation of the viral regulatory elements is a well known mechanism for gene regulation. A recent analysis of the FAAH core promoter (100bp upstream) in human T-cells identified a C/EBP site which (through a STAT3 tethering) mediated the leptin regulation of FAAH expression[MDRFAR03]. Surprisingly, the leptin mediated regulation of FAAH was observed in immune cells, but not a model of neuronal cells[MGF+04]. Our results suggest an alternative regulatory
Figure 3.3: FAAH locus association. The \( x \)-coordinate of each point corresponds to a single 5Kbp locus starting at that location. The \( y \)-coordinate coresponds a permutation based \( p \)-value. The most significant locus is \( \sim 1 \)Kbp upstream of the FAAH transcription start site. The region is part of an LTR element, which are known to carry regulatory signals.

region 1Kbp upstream. An scan for Transcription factor binding sites reveals many interesting sites, including one for C/EBP (data not shown).

The enzyme monoacylglycerol lipase (MGLL), encoded by the MGLL gene located on chromosome 3q21.3, is a presynaptic enzyme responsible for the hydrolysis of 2-arachidonoylglycerol (2-AG), the most abundant endocannabinoid found in the brain. The RARECOVER scan on 923 RVs is shown in Figure 3.4. Here as well, our analysis identifies a single window upstream of the gene, suggesting that the causal RVs have a regulatory function. At the most significant locus (chr3:129030872-129035532, upstream of MGLL TSS), 12 of 25 RVs were selected, with the union RV present in 39 cases, and 8 controls (permutation \( p \)-value \( 6.5 \cdot 10^{-4} \)). The locus contains a known LINE element and a promoter for RNA polymerase II. Mutations in this promoter could easily interfere with binding affinity for RNA Polymerase II and subsequent transcription/translation.
3.3 Causal RVs and Metabolite expression levels

Recently, Sipe and colleagues collected metabolite expression levels on 8 metabolites from 147 individuals from the obesity data-set. Comparing against our FAAH data, these individuals correspond to 14 case individuals with at least one RV in C (identified by RareCover), 84 case individuals with no causal RV, and 49 control individuals. As FAAH helps metabolize AEA (anandamide), and increased levels of AEA are associated with obesity, we expect AEA levels to be higher in cases, especially in the presence of RVs. Figure 3.5 shows that this is indeed the case, with mean values of (17.11 ± 5.58, 15.37 ± 5.45, and 13.71 ± 5.42). A t-test on cases (RV) and controls gives a one-tailed p-value of 0.022. Our results are consistent with the hypothesis that the RVs identified by RareCover influence FAAH gene expression. A similar comparison of MGLL genotypes against 2-AG expression levels did not show significant association, suggesting that other enzymes might be involved in 2-AG metabolism.
Figure 3.4: MGLL Locus association. The most significant locus lies approximately 6Kbp upstream of the MGLL gene.

Figure 3.5: FAAH genotypes and AEA (anadamide) expression levels. The mean expression levels of AEA is significantly higher in cases containing RVs selected by RARECOVER, as compared to controls.
Chapter 4

Discussion

We describe here a novel method for Rare variant analysis that is shown to be competitive against other methods, and helps generate plausible hypothesis on the genetic role of FAAH, and MGLL in models of obesity. The power of the method is parametrized against locus PAR, and penetrance, based on difference in properties of rare variant versus common variants. An interesting observation is that a method based on selecting a combinatorial subset of SNPs outperforms other methods which weight SNPs based on prior observations. This is based on the observation that most causal RVs have functional significance, and likely to have moderately high penetrance, which one would not expect in a non-causal RVs. On the other hand, small weights on a number of non-causal RVs can dilute the association of the causal RVs.

Nevertheless, our study also raises many methodological questions. First, we work with a simple Union model in which the penetrance does not change upon inclusion of additional SNPs, but the PAR increases. Other, more complex models are possible, and will be the focus of subsequent investigation. Our definition of a locus is set arbitrarily as a window of fixed length. It is possible that a dynamic assignment of the size of the locus could increase power, but at the cost of additional computations.

As RVs do not typically show Linkage Disequilibrium with a causal variant, it has been suggested that rare variant analysis leads to the discovery of causal variation. Our results suggest that this is not always true, and even methods that
directly hunt for the subset of causal RVs do not always find them. In our analysis of FAAH, and MGLL, we identified regions that were enriched in RVs that co-segregated with obesity phenotype, and suggested a regulatory mechanism, without ascribing function to the specific subset of RVs. However, it was encouraging that in the two cases sampled, a small region was revealed to be significant.

In this manuscript, we analyze rare and common variants separately. While the RareCover algorithm can work unchanged, with rare and common variants, it is hard to simulate a locus with both rare and common variants affecting disease outcome. Interestingly, the locus identified by us for MGLL also harbors common variants that are significantly associated. A combined statistic would be worth exploring.
Bibliography


