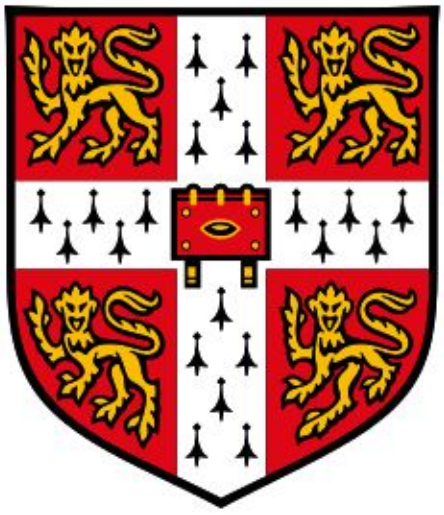
Coverage Analysis in Fine Mapping Genetic Associations

Hope M. Watson

A thesis presented for the degree of
Master of Philosophy



Department of Public Health University of Cambridge
August 2018

Abstract

Acknowledgements

I would like to thank the many people who made this thesis possible. Without their help I would not have grown as an academic and begin to change how I see phenomena occurring in the world. I especially like to thank Dr. Chris Wallace and Dr. Jennifer Asimit, for continuously helping me all throughout this thesis and assisting me in the many questions I asked. I would like to thank Dr. Deborah Thompson for helping edit this thesis and being a devoted supervisor and mentor during my entire year, both academically and personally.

Dedication

For my Dad. Without your love, support, and encouragement I wouldn't be half the person I am today.

Declaration

I declare that this thesis is the result of my own work. All sources are fully acknowledged and appropriately cited. The contents of this submission have not been used to any extent for a comparable purpose, i.e., in partial fulfillment of another degree or diploma at any university or institute of higher learning.

Contents

List of Figures

Chapter 1

Introduction

Genetic testing is increasing at rapid rates, as the cost of genotyping continues to become much cheaper. Genome Wide Association Studies began around 15 years ago. Many strong genetic associations with simple heritability mechanisms have already been identified, while genetic associations with complex genetic mechanisms and disease aetiology continue to be challenging. Determining which associations, that is which mutations to investigate is both time and cost intensive. The overall aim of this study is to assess Bayesian fine mapping techniques. By doing so the goal is to reduce the number of candidate causal SNPs to be considered out of a credible set of SNPs. This is done by improving the current methods of how the credible set is derived. This process enables statistical associations to be converted into smaller sets of candidate target genes.

Chapter 2

Background

The introduction is split into two sections of background to provide clarity from 1) scientific side - genetics, and 2) mathematics side - Bayesian statistics, described in light of the genetics information.

2.1 Genetics

2.1.1 What are Genetic Associations?

A genetic association is a change in the genetic code that alters the risk of disease. Changes occur in the form of single nucleotide polymorphisms (SNPs), copy number variants (CNV), and indels (micro and microsatellites) Genetic association studies -¿ find associations.

2.1.2 What is GWAS?

Genome wide association studies identify genetic positions (loci) associated with different disease risks in a population.

2.1.3 Important Terms - Genetics

A **phenotype** is an expressed trait of interest. This is typically a disease such as cancer or heart disease. In a case-control study, a person with the expressed phenotype of interest is a case, and the person without the expressed phenotype is a control.

A **genotype** is the genetic makeup of a organism that will determine its phenotype. The genetic make up is a sequence of base pairs.

A **single nucleotide polymorphism (SNP)** is a mutation where one base pair is different than the reference genome. Most SNPs are bi-allelic; this makes them easy to be analysed through a binary statistical inferences.

A **gene** a stretch of DNA that codes for a particular protein. A haplotype is a group of genes within an organism that was inherited together from a single parent. Haplotypes also refer to the inheritance of clusters of single nucleotide polymorphisms (SNPs), which are variations at single positions in the DNA sequence among individuals [https://www.nature.com/scitable/definition/haplotype-

haplotypes-142]. Haplotypes are necessary in creating linkage disequilibrium matrices (discussed below).

An **allele** is alternative forms of a gene/trait that are found at the same place on a chromosome. **Minor Allele Frequency (MAF)** is the frequency (count) at which the second most common allele occurs in a given population. The MAF gives a reference to how common or rare a variant is. A MAF of 0.5 means that the frequency of that specific allele is present in 50 percent of the population.

Understanding GWAS Manhattan Plot

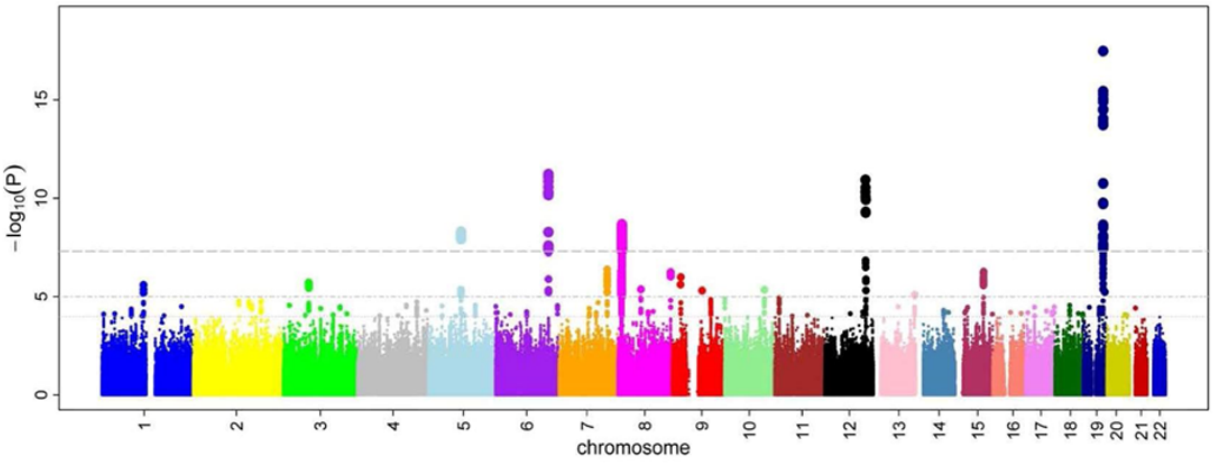


Figure 2.1: Manhattan Plot

Each dot on a Manhattan plot represents one haplotype. The x axis is the loci, which is the position of the SNP on the chromosome. On the y axis is the -log p value of the association of the SNP in the cases compared to the control group. Dots farther up on the y axis represent a SNP that has a stronger association with the phenotype. The null hypothesis is that each SNP has no association with the phenotype. Through GWAS, candidate genes can be identified. In plots where there are multiple highly associated SNPs, this is an indication of pleitropy, which means that the same gene effects more than one phenotype.

2.1.4 What is Linkage Disequilibrium?

Linkage disequilibrium is the correlation structure that exists among DNA variants in the current human genome as a result of historical evolutionary forces, particularly finite population size (genetic drift), mutation, recombination rate, and natural selection [7]. In short, this correlation structure means that SNPs are not statistically independent of each other. LD makes it possible for 80 million SNPs to be arrayed by only 500kb-1Mb SNPs. However, LD also makes it difficult to indicate which SNP is causal and which is a highly correlated neighbouring SNP. Different genes have different recombination probabilities, and therefore genetic distance does not equal physical distance.

LD is expressed by a matrix, where

2.1.5 Fine Mapping

Fine mapping is a process in which researchers try and identify the most likely single or set of variants within which the causal variant should be present. The process is done by assigning well-calibrate probabilities of causality to each candidate variant [5].

Types of Methods in Fine Mapping

Lead SNP (smallest p-value) A previous and simple approach was to consider all SNPs of a certain threshold (5 x 10^{-8}) as potential candidates for causality. There are several issues with this, as p-values are influenced by study specific factors such as power (sample size), minor allele frequency, and the effect size (rarely known). Differences in these variables make genetic studies difficult to compare [5].

Lead SNP and LD "friends" (i.e. r^2>0.8) Another method used was to take the lead SNP, that is the SNP with the highest p-value, and compare it with a certain LD threshold, typically r^2 >0.8) and consider these neighboring SNPs as the potential causal SNP. This method still ignores properties of the study or locus, as greater power can differentiate SNPs in higher LD.

Bayesian method Association p-values converted to Bayesian posterior probabilities of causality under specific assumptions [6] [8]. These can be summed over sets of variants to generate a posterior probability a causal variant lies in any given set. The posterior probabilities (pp) are the ratio of evidence for each variant being causal versus all the others. This is expressed as

$$\underline{pp \sum}$$

equation

2.1.6 Posterior Probability

The Bayesian's belief in a binary hypothesis (eg this SNP is causal vs this SNP is not causal) after seeing the data. Note difference to prior belief. Bayes Theoreom

$$P(X = x|D) = \frac{P(X = x)P(D|X = x)}{P(D)} \tag{2.2}$$

$$P(X = x|D) = \frac{P(D|X = x)P(X = x)}{\sum(P(D|X = y)(P(X = y))} \tag{2.3}$$

$$P(X = x|D) \propto P(D|X = x)P(X = x) \tag{2.4}$$

statistical probability that a hypothesis is true, calculated in light of relevant observations. Relevant observations may be defined as both the prior information and the new data that is being analysed, both which together generate the posterior probability. For this reason, the posterior probability is proportionally related to the prior, by the expression of the newly observed data.

Relationship between p-values and Posterior Probabilities The relationship between p-values and posterior probabilities is an inverse relationship. The smaller the p-value which represents the likelihood of observing the data given the null hypothesis, the higher the posterior probability.

- typically, the SNP is not associated with the

2.2 Bayesian Statistics

2.2.1 Contrast with Frequentist

In a frequentist framework a parameter of interest (mean, proportion, rate) is considered fixed and only its estimates vary due to sampling variation. The process of inferring the range of values of parameter works by considering every possible result that a study could potentially generate. That is, under the same conditions, what would be observed under multiple parallel samples. However this is not possible for many biological phenomena, where a question such as 'What is the probability that a SNP is causal, given the data currently accrued?' [@Kirkwood2005]

In frequentist statistics the null hypothesis is assumed to be true; that is, it is the starting point of all inference that the observed data is compared to. The p-value reflects the probability of observing data as or more extreme than the data actually observed in the current study, given the null hypothesis [2]. A 95% confidence interval is interpreted as, if new data were to be repeatedly sampled and analysed, 19 out of 20 intervals would include the true quantity parameter of the being estimated.

Bayesian statistics incorporate pre-existing information into analyses, in the form of a prior, which is a belief of the distribution of the data before any new data has been observed or incorporated in analysis. This can come a variety of forms such as previous studies, consultation with experts, or theoretical biological models. Under these conditions, spelt out by the Bayes Theorem, a parameter is correctly interpreted as the probability of a hypothesis , given the observed data.

2.2.2 Link with Frequentist

Here, we discuss the link between a frequentist confidence interval and a Bayesian credible interval with an uninformative prior. When the Bayesian framework is used with an uninformative, the posterior probability is derived from the observed data. The definition of likelihood is bas. An uninformative prior mathematically takes the form of uniform distribution, so that any probability in the distribution has the same proportion. This is similar to frequentist probabilities, which only use the data to estimate the unknown parameter.

2.2.3 Important Terms - Statistics

Credible interval A credible interval is a range of values within which an unobserved parameter values falls within a particular subjective probability [Edwards1963].

Credible set A set of SNPs that contains the causal variant with a pre-specified probability. Each SNP has an assigned prior probability and a posterior probability is calculated, using the likelihood of the data observed [Cite Jenna somewhere here].

The set is determined by a threshold of desired cumulative poster probabilities to reach this cutoff.

Issues with Credible Sets There is currently no explicit quantitative definition for a credible set as there is for a credible interval. A credible assumes a continuous distribution, while a credible set is made up of a discrete distribution for the SNP

Size of credible set For one credible set, the size of the credible set is the belief that the credible set contains the causal variant (eg 90

Coverage of credible set If we repeat analysing a list of credible sets, coverage is the amount of times the causal variant is in that credible set. Here coverage can be defined as number of times the causal variant was in the credible set out of the number of simulations run. This can be expressed as a percentage where if 10 credible sets are analysed and 9 of these credible sets include the causal variant, there is 90% coverage. *Coverage is a frequentist concept because of the nature of repeatability being used to assess Bayesian inferences.*

2.3 Scope of the Study

This study is done under the one causal variant assumption, meaning there is only one causal variant per credible set. This assumption creates some limitations that are outlined here. Colocalization is when a single SNP is associated with multiple phenotypes. Some methods used for colocalization calculate posterior probabilities works by fine mapping each trait under a single causal variant assumption. The two posterior probabilites are then integrated over to calculate probabilities that those variants are shared [@Fortune2018]. Therefore, this analysis may be able to be extended to involve colocalization in the future but is not addressed at this time.

Copy number variants (CNVs) and variable number tandem repeat (VNTRs) consisting of microsatellites and minisatellites mutations, may account for a 10-15% of heritable gene expression variation in humans [@Gymrek2016]. Repetitive DNA is not easily analysed by next generation DNA sequencing methods, which struggle with homopolymeric tracts, that is, parts of the genotype that have the same base pair type repeated many times. Bi-allelic SNPs are not subject to this issue in the way CNVs and VNTRs are, which means they are more easily assayed. This analysis does not address the potential importance of CNVs and VNTRs and assumes that the causal variant can is a bi-allelic SNP.

Functional annotations are added information known about a SNP in how it effects expressed phenotypes. Functional annotations are typically expressed as coding or non-coding regions of the genes. Among functional annotations are expression quantitative trait loci (eQTL) where a SNP is known to effect a level of expression a particular gene in a particular tissue. Functional annotations are important because phyiscal distance to a gene is not substantive evidence of causality. This means a mutation farther from a gene may play an important role in regulation of that specific gene's expression. Although not explored here, methods outlined in this study could be explored by altering the assigned prior value of a SNP in finemap.abf. Reweight-

ing of the posterior probabilities can be done by using fGWAS [4] or PAINTOR methods [3].

Different trans-ancestries have different LD patterns. This study does not address any trans-ancestry differences or geographical analysis of subpopulations and their respective potential subtypes. A SNP that has a small p-value across groups that have different genetic architecture shows a stronger genetic association to a phenotype than if it were present in only one of the groups. In this analysis only one LD matrix is considered at a time, but this could be extended in the future, particularly with relevant open data access from sources such as 1000 Genomes Project. [find spot to cite Jenna's papers here].

2.4 Aims

This study aims to assess Bayesian fine mapping techniques. The study does this by analysing current fine mapping practices and considering how they may be improved. To do this, a metric to capture the amount of disorder in the system - the set of SNPs, was created. Through this method, we are able to say that by making the SNPs with the highest posterior probabilities always incorporated in the credible set through ordering, this contains more information about each SNP. By incorporating the information that SNPs with high posterior probabilities are more likely to be causal, the specified threshold can be reached using a smaller credible set.

Should go somewhere else soon, but in the flow

The measure of disorder can be viewed as a statistical annotation that allows us to have more confidence that the causal variant is contained in the credible set, when candidate SNP(s) have much higher posterior probabilities than the entire full set of SNPs being tested.

2.5 Current Practices in Fine Mapping

Currently it is typical practice that credible sets be ordered by descending posterior probabilities to order the SNP with the highest posterior probability first and the SNP with the lowest posterior probability last. This practice is done because it reduces the credible set, that is the number of candidate SNPs to consider.

Under this practice, the relationship to size and coverage are considered.

2.5.1 Relationships to Ordering

Size The relationship between ordering and credible set size is an inverse relationship. This is simple to see in an example. Say there are 100 SNPs and the 97th SNP is the causal variant with the lowest p-value, and therefore has the highest posterior probability. An unordered credible set with a pre-specified threshold (cumulative sum of posterior probabilities), would need to take 97 SNPs to achieve the specified threshold. The ordered credible set could achieve the same threshold with many fewer SNPs. This is illustrated below.

Coverage

Chapter 3

Methods

3.1 Simulations

3.1.1 Why use Simulations?

Simulations are the gold standard in evaluating methodologies relevant to causal variants. In the "real-world", that is genotyping for actual phenotypes of interest (typically a disease), we do not know the causal variant. It is not possible evaluate both real-world data and the robustness of the methodology simultaneously. This is because the lack of certainty and variability of both parameters may be the reason for the outcomes; a basic concept of the scientific method – test one variable at a time and hold all other variables the same as controls.

3.1.2 Running Simulations

Simulations were run by through a series of steps, leveraging relevant available R packages. Full code described here is available on: https://github.com/HopeMWatson/credibleset$_t$h

The libraries utilised in this analysis were devtools to load the simGWAS and coloc packages directly from github. ggplot2 package was used to produce figures.

3.2 Functions and Packages

3.2.1 simGWAS - Generating p-values

The simGWAS package directly simulates GWAS summary data, without individual data as an intermediate step. The expected statistics are mathematically derived for any set of causal variants and their effects sizes, conditional upon control haplotype frequencies [1]. The arguments for simGWAS to run require a specification of the causal SNP, its effect size (OR), and allele (MAF) and haplotypes frequencies. simGWAS then produce simulated z-scores for each SNP. A list of z-scores was created for each SNP by running the simulation 100 times. The final output from simGWAS was a 100x100 matrix, which contained data on 100 SNPs with 100 z-scores. The z-scores were turned into p-values.

3.2.2 Finemap.abf - Generating posterior probabilities

From the assigned p-values, posterior probabilities are calculated. This process was outlined above, showing how approximate bayes factors (ABF) are created from the observed data and the uninformative prior. The finemap.abf function works on summary statistic data, inputs as either 1) p-values, as used here or 2) coefficients and coefficient variance. Other functions in the coloc package can be utilised for full genotyped data with different statistical approaches.

The finemap.abf function works by specifying input arguments for 1) p-values 2) sample size 3) MAF 4) ratio of cases to controls (s), and 5) type =cc. The prior is set to p=1e-04 as a default for an uninformative prior. An empty matrix was created in to map each to place each posterior probability into the list back into the data frame. The finemap.abf function was then looped over each SNP and its respective list of p-values to create posterior probabilities. The final output from finemap.abf was a 100x100 matrix, which contained data on 100 SNPs with 100 respective posterior probabilities of the SNP being the specified causal variant.

3.2.3 Credset - Generating Credible Set

The credset function produced a credible set based upon the desired threshold. The threshold represents the *cumulative* posterior probabilities *in the original sorting generated* to reach or *exceed* the threshold value. In the cases of exceeding the threshold, this is because the a SNP's posterior probability is necessary to reach the threshold, but large enough to exceed it, where "overshooting" occurs.

Credset analyses both ordered and not ordered methods. The ordered methods first sort by descending posterior probabilities, then takes the cumulative probability to reach the threshold. Not ordered keeps the original sorting of how the posterior probabilities were generated and uses the seq_along function to take the all posterior probabilities of the SNPs until the threshold is reached. This returns the credible set - a set of candidate SNPs contains the causal variant with a pre-specified probability.

Credset produces a credible set for *one simulation*. The cumulative posterior probabilities is the size. There is no interpretation of coverage for one credible set, since there is only one simulation and coverage is by definition analysis under a repeated process. However, the variable **contained** is defined as if the causal variant was in the the credible set. For one *repetition* (nrep=1) of one simulation contained takes on the value of 0 - the causal variant was not in the credible set or 1 - the causal variant is in the credible set. If nrep=10, the possible values for one reptition of the simulation would be 0.1 if causal variant is in the credible set or 0 if not.

Contained is related but markedly different from coverage. Contained still only addresses one credible set, even if the simulation for that same credible set is run x number of times (nrep=x). Coverage addresses a number of credible sets. Coverage is calculated by taking the values for contained over the list of credible sets analysed.

3.2.4 Wrapper - Size and Coverage

While credset analyses one simulation, the wrapper function applies the credset across multiple simulations. This returns the probabilities - average size and coverage of a list of credible sets.

Size and coverage is reported for both ordered and not ordered sets.

3.2.5 Quantifying Disorder

Disorder was mathematically defined as:

$$disorder = -sum(log(pp)) \tag{3.1}$$

Disorder is not ordering dependent, that is the disorder of a system is constant for the determined set of credible SNPs 3.2.5 explores this further.

The mathematical term for entropy was *not* chosen due to the extraneous values that approach infinity and negative infinity when multiplying values over a distribution. Where $entropy = -sum(pp * log(pp))$.

Visualising Disorder in a Credible Set

To demonstrate how disorder varies in a credible set, we give a visual example with 10 hypothetical SNPs.

In this first diagram there is no disorder in the system. It is a uniform distribution, where all SNPs share the same posterior probabilities. Each SNP is equally likely (or unlikely) to be the causal variant.

The second diagram shows medium disorder where there is a SNP with a large posterior probability - the lead SNP, in addition to SNPs in high LD with the lead SNP. Determining which of the candidate SNPs is the true causal variant is difficult, as multiple SNPs appear good candidates. This is the scenario that occurs almost always in real world data, and is the focus of fine mapping.

The third diagram shows a system with very high disorder, where the lead SNP accounts for almost all of the cumulative posterior probability. In this situation we can safely infer this is *the* causal variant. Most associations with this degree of strength have already biologically defined and well studied.

Finally, we see if we mixed up the order of the SNPs, it does not change the amount of disorder captured.

3.2.6 Wrapper2

Wrapper2 generates a credible sets under different conditions of odds ratio (OR) and sample size number (n), with a specified threshold. The credible set produced can take on any combination of OR=1.0, 1.05, 1.1, 1.2, 1.3 and the n=1000-5000 in increments of 1000. The simultaneous combinations allow analysis of how the property of disorder effects the number of variants (nvar) in a credible set. This is explored in the results section 4.3.

3.2.7 Covent

Covent is both the function that replicates wrapper2 to create a loop to generate more simulations and is the data produced from the simulations. a view of the data can be found in 4.1.

For the specified number of replications, the data created are replicated for *both* cases and controls. Therefore, if 1000 replications are specified, there will be 2000 total simulations, 1000 case simulations and 1000 control simulations.

3.2.8 Logistic Regression Testing

Significance testing for disorder was conducted using logistic regression and compared in order versus not ordered sets.

Chapter 4

Results

4.1 Different Scenarios

Different size and coverage scenarios were explored to analyse the relationship ordering.

Each set of simulations had a set seed (set.seed(42)) to generate reproducible results. Two hundred simulations were run for each scenario. The same 200 simulations were produced for OR and threshold due to the seed. Each simulation had 10 repetitions. There was no set seed for sample size scenarios to allow for variability.

Each scenario used a default threshold of 0.9 and odds ratio of 1.5. This is to follow the basic scientific principle to only change one variable at a time while holding the rest as controls. All covariate values are explicitly specified for each set of simulations for clarity and confidence intervals are recorded as 95% CIs.

Scales have been adjusted per each simulation on purpose, so trends may be easily identified. All specific values from simulation outputs can be found in tables in the Appendix.

4.1.1 Ordered vs. Not Ordered

Analysis of how ordering effects size was considered. A threshold of 0.9 and odds ratio of 1.5 was used. Simulation results support fundamental and intuitive understanding that ordering reduces the credible set size. Ordered sets had on average 4.4% smaller size (3.7% to 5.2%).

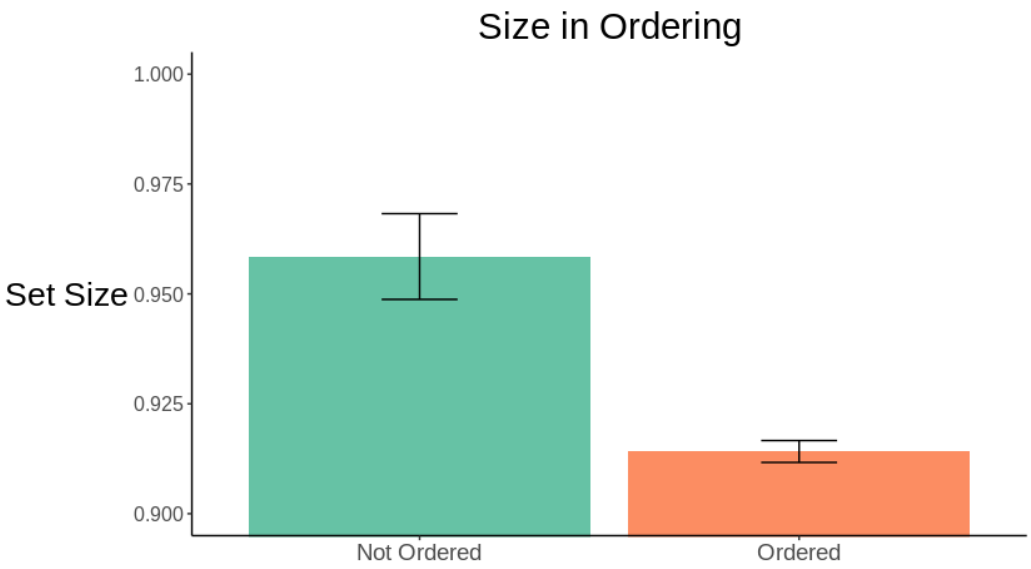


Figure 4.1: Ordered vs. Not Ordered Size

Coverage is higher in ordered sets compared to not ordered sets, after controlling for odds ratio (1.5), threshold (0.9), and sample size N cases(1000) and N controls (1000). The simulation proves that coverage is effected by ordering. This means that when data are ordered, it introduces information about containing the causal variant, outside of the sum of posterior probabilities. We will refer to this as overcoverage, where the specified value for a threshold is exceeded due to additional information introduced from ordering the sets. Coverage was 4% greater in ordered sets (4.8% to 3.2%).

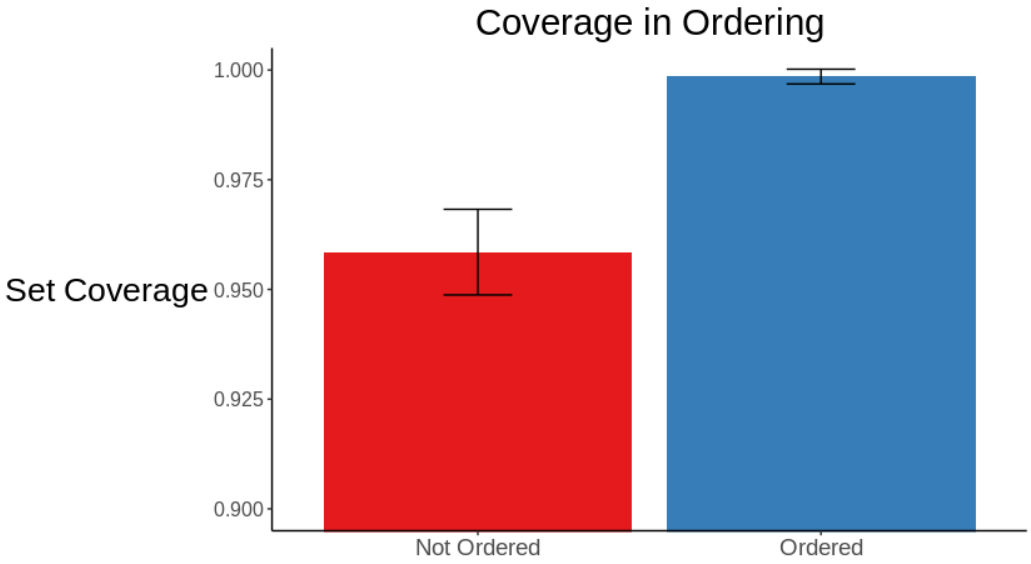


Figure 4.2: Ordered vs. Not Ordered Coverage

Trends were analysed across different covariates that effect size and coverage of sets.

4.1.2 Odds Ratios

Odds ratios represent the effect size (Θ) that a SNP has on the association of the phenotype. The larger the odds ratio, the stronger the signal, and therefore more confident inference for which variant is causal. Odds ratios were explored within a sensible range of 1.05 to 1.5. For all odds ratios simulations, the threshold was kept at 0.9.

Across all odds ratio values the average size of the credible set was reduced when ordering. Sets that were not ordered, were much larger for OR = 1.5. This is because the signal for that variant is strong, and has a high posterior probability.

WHY IS 1.5 SO MUCH GREATER AND NOT INCREMENTALLY GREATER?

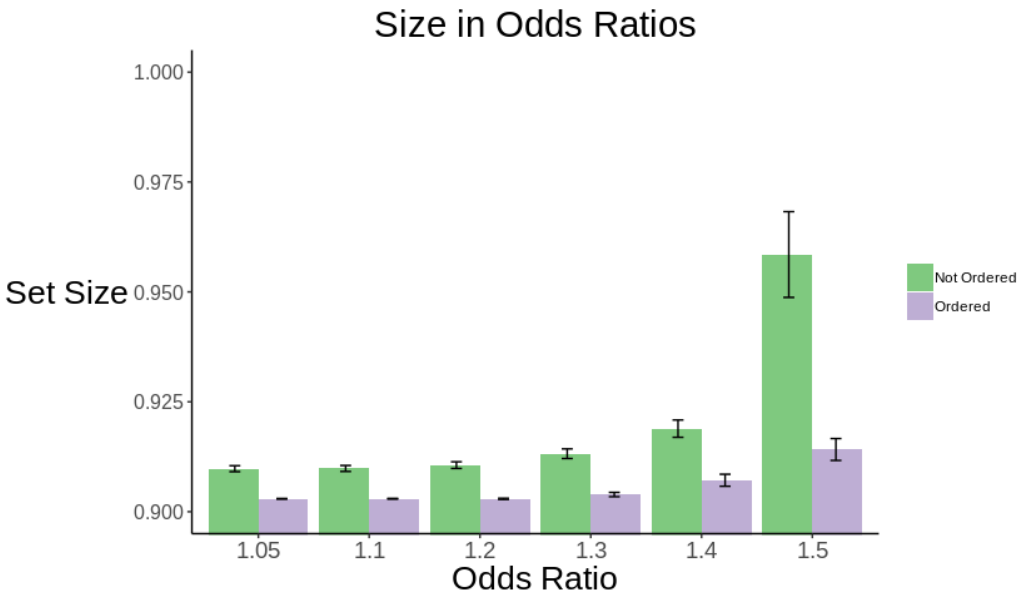


Figure 4.3: Odds Ratios and Size

WAS NOT EXPECTING THESE RESULTS OF INCONSISTENCY - ASK ABOUT.

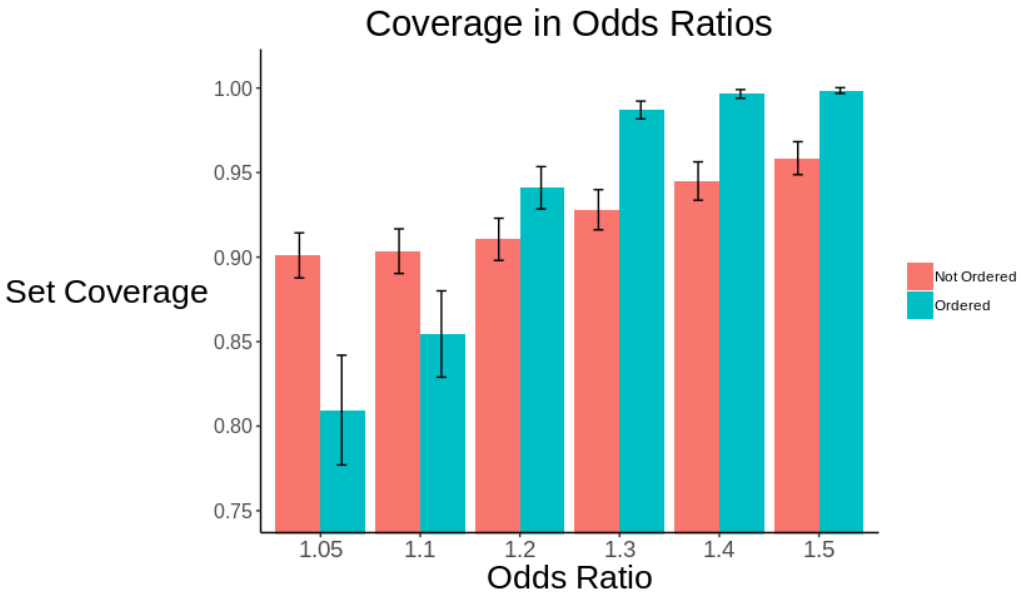


Figure 4.4: Odds Ratios and Size

4.1.3 Thresholds

Different threshold levels were investigated where threshold was set at 0.5, 0.9, and 0.99. Controlled variables were OR (1.5) and sample size N cases (1000) and N controls.

Ordering reduced set sizes across all thresholds. Average difference in size and variance is greater in lower thresholds than higher thresholds.

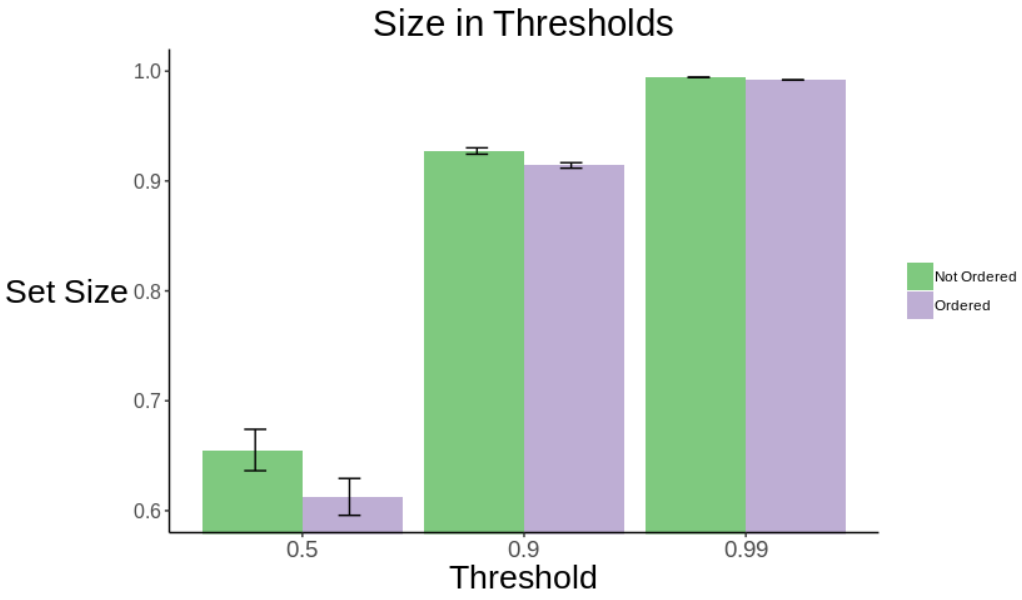


Figure 4.5: Thresholds and Size

Coverage was analysed across different thresholds. Simulations revealed overcoverage in ordered sets at the 0.5 and 0.9 level. At the 0.9 level the ordered coverage upper confidence interval was marginally over one, an unreal result. However, at

the 0.99 level the not ordered coverage upper confidence interval returned an unreal result over one.

Overcoverage for threshold=0.5 was 23.3% (24.8% to 21.8%); threshold 0.9= 4.0% (4.8% to 3.2%). threshold=0.99 0.05% (1.5% to -0.0048%)

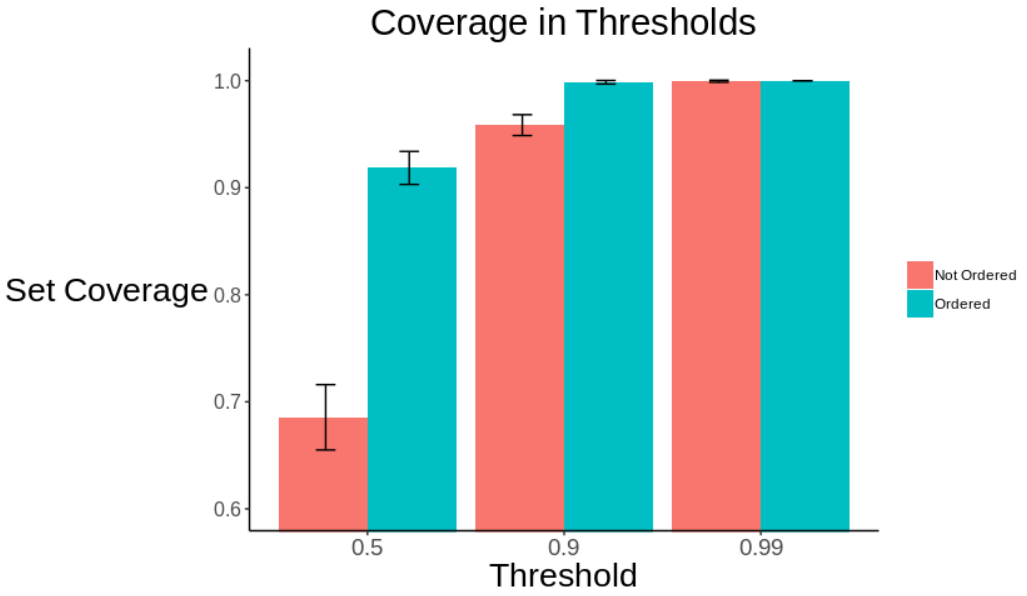


Figure 4.6: Thresholds and Coverage

4.1.4 Sampling Size (n)

Different scenarios of sampling were simulated. The ratio of cases and controls was kept the same where s=0.5 throughout the different sampling numbers tested. The sample sizes tested were 1000, 5000, 10000. There was no seed used in sample size simulations to allow for variability. For this reason, the sample size of 5000 has more variability - not really sure about this.

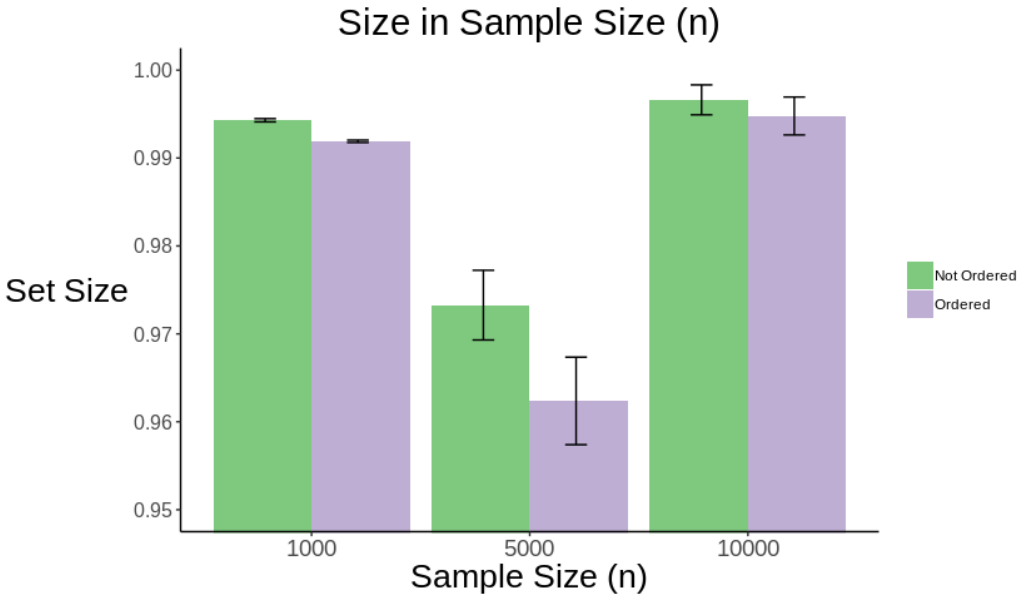


Figure 4.7: Sample Size (n) and Size

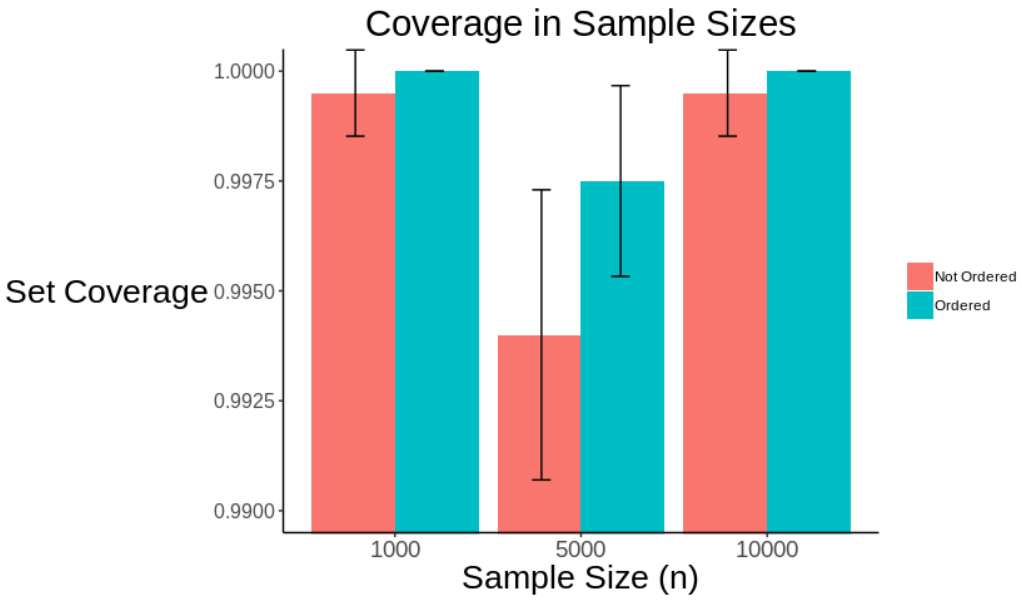


Figure 4.8: Sample Size (n) and Coverage

4.2 Relationships to Disorder

We simulated scenarios with varying values for threshold, sample size, and odds ratio. This was done using the code described in 3.2.6. The size and the coverage of the credible set is recorded for each simulation. 1000 credible sets were created, example data from is shown below. These data were used to analyse the relationships between a) -sum(log(pp)) and size b) -sum(log(pp)) and coverage.

Simulations were run 100, 500, and 1000 times comparing graphs and outputs. Logistic regression outputs for 1000 simulations are stated in this chapter, while regression outputs for 100 and 500 simulations can be found in the appendix.

Table 4.1: Sample Input Data

Ordered	Thr	Size	Number of Variants	Covered	Disorder	N	OR
TRUE	0.500	0.508	29	0	92.758	5,000	1.050
TRUE	0.500	0.799	1	1	133.538	2,000	1.300
TRUE	0.500	0.506	15	1	104.670	4,000	1.100
TRUE	0.500	0.504	21	1	108.682	5,000	1
TRUE	0.500	0.509	19	0	115.005	5,000	1.100

4.3 Size and System Disorder

The relationship between credible set size and disorder was explored. In we can see that in the not ordered sets the number of variants in the credible set, size, is not effectively reduced. Disorder remains constant whether the sets are ordered or not ordered. We can see proof of this, as the disorder coefficient and variance is similar at 100 simulations and the same by 500 simulations.

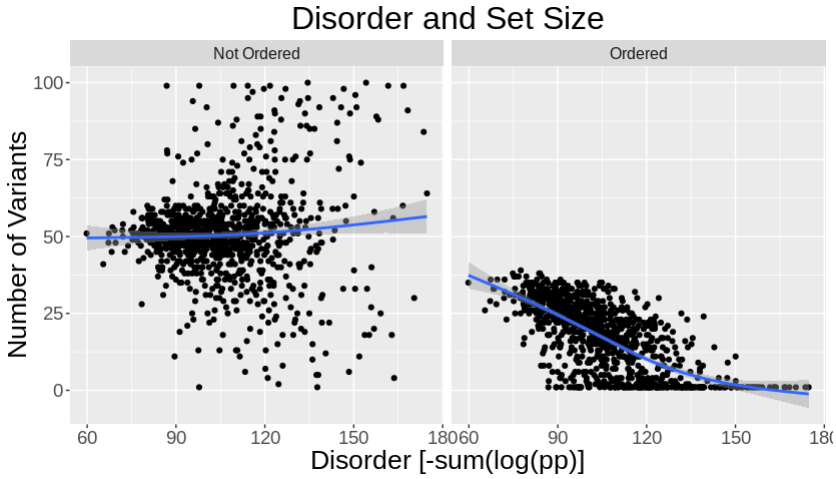


Figure 4.9: 100 Simulations

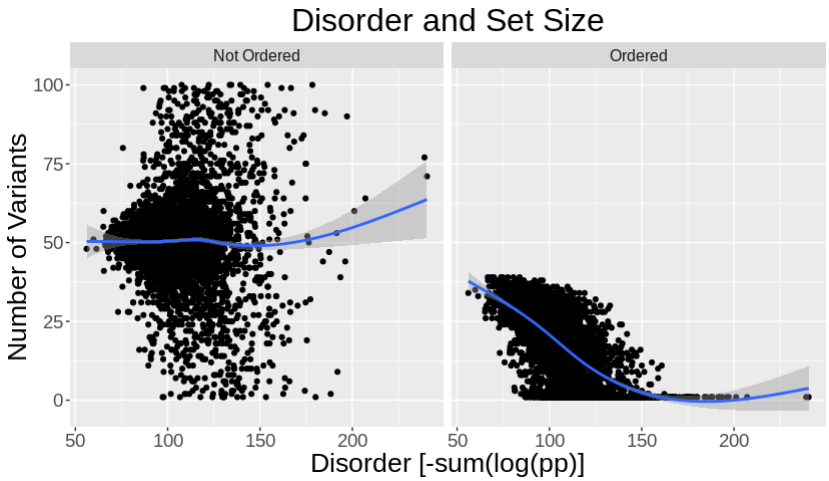


Figure 4.10: 500 Simulations

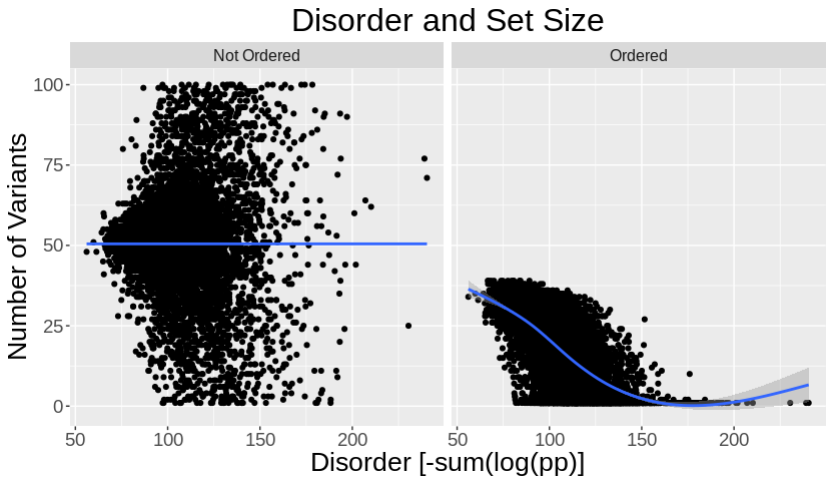


Figure 4.11: 1000 Simulations

The difference is how the trend changes in ordered versus not ordered sets. In not ordered sets, there is no general trend - indicated by the flat regression line.

For the ordered sets, there is a downward trend, where as entropy increases, the size of the set gets smaller. This trend occurs because as entropy increases in ordered sets, we are increasing the the number of small p values - high association signals. This means we are also increasing the number of large posterior probabilities of the SNP being the causal variant. Very high disorder systems, as explored in 3.2.5 have smaller sized sets, because the posterior probability one SNP is far greater than the other SNPs. Overall, the number of variants becomes smaller because we are considering less SNPs to be the causal variant.

4.4 Coverage and System Disorder

Here we are showing that as -sum(log(pp)) increases, the coverage increases.

The analysis shows that -sum(log(pp) is important and useful to capture, because it helps improve coverage – we are better able to make accurate claims for coverage of specified intervals.

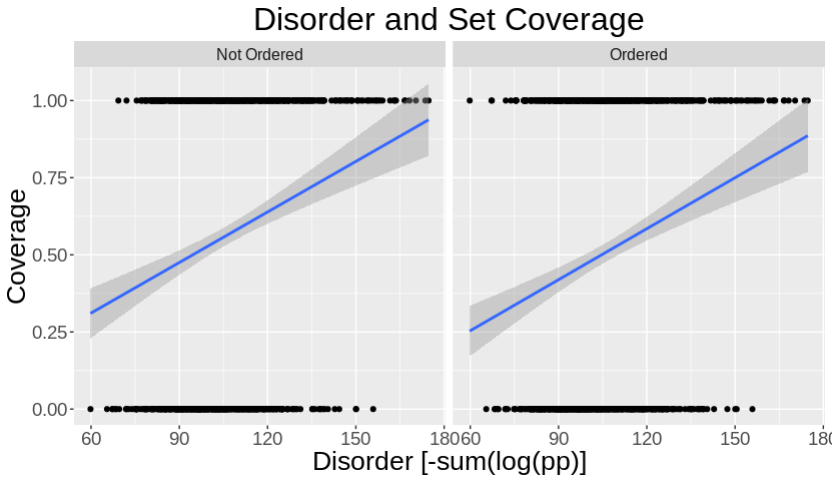


Figure 4.12: Coverage and Disorder in 100 Simulations

Table 4.2: Ordered Sets: Coverage vs. Disorder Logistic Regression

((a)) 100 simulations

Coefficient	Estimate	Std. Error	z value	P-value
Intercept	-2.258	0.402	-5.13	$1.98e-08$
Disorder	0.0238	0.0038	6.26	$3.83e-10$

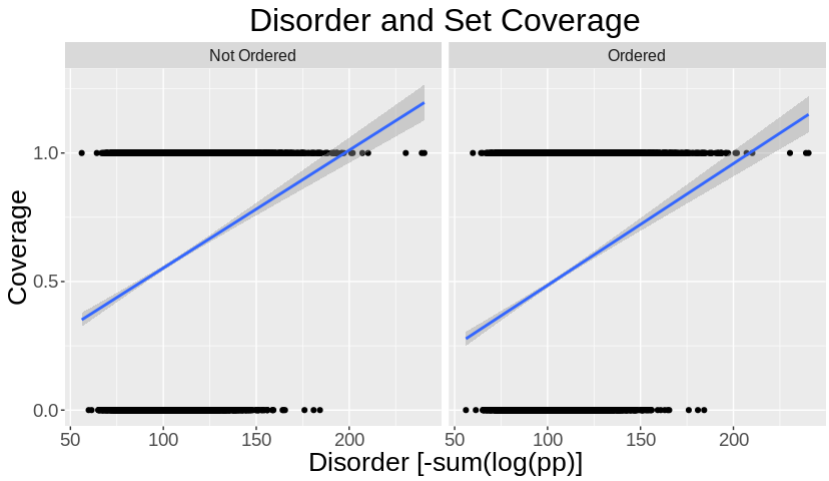


Figure 4.13: Coverage and Disorder in 1000 Simulations

4.5 Relationship between Inherent Property of Disorder and SNPs

The relationship between disorder

4.5.1 Set Seed Regression

4.5.2 No Seed Regression

4.6 Creating a Correction Factor for Credible Set

fix!

Chapter 5

Discussion

5.0.1 Property of Disorder

The LD matrix in the simulation was artificial, in that the points of high LD "blocks" and low LD "hotspots" were very symmetric. This is not found in actual parts of the genome. Therefore the behaviour between disorder and number of SNPs is not as perfectly linear. This was proven by running the analysis without a seed.

Chapter 6

Conclusion

Appendix A

Appendix

Table A.1: Ordered Size

	Ordered.Unordered	Size.Value	LCI	UCI
1	Ordered	0.914	0.912	0.917
2	Not Ordered	0.958	0.949	0.968

Table A.2: Ordered Coverage

	Ordered.Unordered	Cov.Value	LCI	UCI
1	Ordered	0.998	0.997	1.000
2	Not Ordered	0.958	0.949	0.968

Table A.3: OR Size

	Ordered.Unordered	OR	Size.Value	LCI	UCI
1	Not Ordered	1.050	0.910	0.909	0.910
2	Not Ordered	1.100	0.910	0.909	0.911
3	Not Ordered	1.200	0.911	0.910	0.911
4	Not Ordered	1.300	0.913	0.912	0.914
5	Not Ordered	1.400	0.919	0.917	0.921
6	Not Ordered	1.500	0.958	0.949	0.968
7	Ordered	1.050	0.903	0.903	0.903
8	Ordered	1.100	0.903	0.903	0.903
9	Ordered	1.200	0.903	0.903	0.903
10	Ordered	1.300	0.904	0.903	0.904
11	Ordered	1.400	0.907	0.906	0.909
12	Ordered	1.500	0.914	0.912	0.917

Table A.4: OR Coverage

	Ordered.Unordered	OR	Cov.Value	LCI	UCI
1	Ordered	1.050	0.810	0.777	0.842
2	Not Ordered	1.050	0.901	0.888	0.914
3	Ordered	1.100	0.854	0.829	0.880
4	Not Ordered	1.100	0.904	0.890	0.917
5	Ordered	1.200	0.941	0.928	0.954
6	Not Ordered	1.200	0.910	0.898	0.923
7	Ordered	1.300	0.987	0.982	0.992
8	Not Ordered	1.300	0.928	0.916	0.940
9	Ordered	1.400	0.996	0.994	0.999
10	Not Ordered	1.400	0.945	0.934	0.956
11	Ordered	1.500	0.998	0.997	1.000
12	Not Ordered	1.500	0.958	0.949	0.968

Table A.5: Threshold Size

	Ordered.Unordered	Sample.Size	Size.Value	LCI	UCI
1	Ordered	1,000	0.992	0.992	0.992
2	Unordered	1,000	0.994	0.994	0.994
3	Ordered	5,000	0.962	0.957	0.967
4	Unordered	5,000	0.973	0.969	0.977
5	Ordered	10,000	0.995	0.993	0.997
6	Unordered	10,000	0.997	0.995	0.998

Table A.6: Threshold Coverage

	Ordered.Unordered	Threshold	Cov.Value	LCI	UCI
1	Ordered	0.500	0.918	0.903	0.934
2	Not Ordered	0.500	0.686	0.655	0.716
3	Ordered	0.900	0.998	0.997	1.000
4	Not Ordered	0.900	0.958	0.949	0.968
5	Ordered	0.990	1	1	1
6	Not Ordered	0.990	1.000	0.999	1.000

Table A.7: Sample Size (n) Size

	Ordered.Unordered	Sample.Size	Size.Value	LCI	UCI
1	Ordered	1,000	0.992	0.992	0.992
2	Not Ordered	1,000	0.994	0.994	0.994
3	Ordered	5,000	0.962	0.957	0.967
4	Not Ordered	5,000	0.973	0.969	0.977
5	Ordered	10,000	0.995	0.993	0.997
6	Not Ordered	10,000	0.997	0.995	0.998

Table A.8: Sample Size Coverage

	Ordered.Unordered	Sample.Size	Cov.Value	LCI	UCI
1	Ordered	1,000	1	1	1
2	Not Ordered	1,000	1.000	0.999	1.000
3	Ordered	5,000	0.998	0.995	1.000
4	Not Ordered	5,000	0.994	0.991	0.997
5	Ordered	10,000	1	1	1
6	Not Ordered	10,000	1.000	0.999	1.000

Table A.9

	Number.of.SNPs	Seeded.Disorder	Unseeded.Disorder
1	100	211,467.300	207,118.600
2	200	409,357.600	407,735.200
3	300	610,161.900	610,014.700
4	400	808,115.700	807,683.500
5	500	1,010,836	1,015,019

Bibliography

[1] Mary D Fortune and Chris Wallace. "simGWAS: a fast method for simulation of large scale case-control GWAS summary statistics". In: *bioRxiv* (May 2018), p. 313023. DOI: 10.1101/313023. URL: https://www.biorxiv.org/content/early/2018/05/02/313023.

[2] L C Gurrin, J J Kurinczuk, and P R Burton. "Bayesian statistics in medical research: an intuitive alternative to conventional data analysis." In: *J. Eval. Clin. Pract.* 6.2 (May 2000), pp. 193–204. ISSN: 1356-1294. URL: http://www.ncbi.nlm.nih.gov/pubmed/10970013.

[3] Gleb Kichaev et al. "Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies". In: *PLoS Genet.* 10.10 (Oct. 2014). Ed. by Anna Di Rienzo, e1004722. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1004722. URL: http://dx.plos.org/10.1371/journal.pgen.1004722.

[4] Joseph K Pickrell. "Joint analysis of functional genomic data and genome-wide association studies of 18 human traits." In: *Am. J. Hum. Genet.* 94.4 (Apr. 2014), pp. 559–73. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2014.03.004. URL: http://www.ncbi.nlm.nih.gov/pubmed/24702953%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3980523.

[5] Sarah L Spain and Jeffrey C Barrett. "Strategies for fine-mapping complex traits." In: *Hum. Mol. Genet.* 24.R1 (Oct. 2015), R111–9. ISSN: 1460-2083. DOI: 10.1093/hmg/ddv260. URL: http://www.ncbi.nlm.nih.gov/pubmed/26157023%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4572002.

[6] Matthew Stephens and David J. Balding. "Bayesian statistical methods for genetic association studies". In: *Nat. Rev. Genet.* 10.10 (Oct. 2009), pp. 681–690. ISSN: 1471-0056. DOI: 10.1038/nrg2615. URL: http://www.ncbi.nlm.nih.gov/pubmed/19763151%20http://www.nature.com/articles/nrg2615.

[7] Peter M Visscher et al. "10 Years of GWAS Discovery: Biology, Function, and Translation." In: *Am. J. Hum. Genet.* 101.1 (July 2017), pp. 5–22. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2017.06.005. URL: http://www.ncbi.nlm.nih.gov/pubmed/28686856%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5501872.

[8] Jon Wakefield. "A Bayesian measure of the probability of false discovery in genetic epidemiology studies." In: *Am. J. Hum. Genet.* 81.2 (Aug. 2007), pp. 208–27. ISSN: 0002-9297. DOI: 10.1086/519024. URL: http://www.ncbi.nlm.nih.gov/pubmed/17668372%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1950810.