Draft thesis 20_07_2018

Author:
Hope Watson

Supervisors:
Dr. Deborah Thompson Dr. Chris Wallace

August 2018

Thesis submitted in partial fulfillment of the degree of Master of Philosophy in Epidemiology

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 fulfilment of another degree or diploma at any university or institute of higher learning.

Word Count (excluding tables, figures, and references):

Table of Contents

Introduction

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), call number variants (CNV), and indels (micro and microsatellites) Genetic association studies -> find associations.

What is GWAS?

Genome wide association studies identify genetic positions associated with different disease risks in a population

What is LD?

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. Visscher2017 In short, this correlation structure means that SNPs are not independent of each other, that is they are conditionally linked to one another. This is what always 80 million SNPs to be arrayed by only 500kb-1Mb SNPs, but does not help indicate which SNP is causal and which is a highly correlated neighboring SNP. Different genes have different recombination probabilities, and therefore genetic distance does not equal physical distance.

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 lie.

Types of Methods

lead SNP (smallest p)

A previous and simple approach was to consider all SNPs of a certain threshold (5 x 10ˆ-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.

lead SNP + LD friends (eg r2>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 rˆ2>0.8 and consider it as a potentially 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. These can be summed over sets of variants to generate a posterior probability a causal variant lies in any given set.

Credible sets - involves sorting

Aims of thesis - assess Bayesian fine mapping - hopefully improve it!

Describing this work requires defining and understanding key terms:

–In frequentest statistics probability comes into play before collecting data. That is alpha level is predetermined and static. 95% probability that we will collect data that produces and interval that contains the true parameter. SAMPLING - MULTIPLE PARALLEL IMAGINARY SAMPLES

–In Bayesian statistics probability comes into play after collecting data. Analysis probability after observing the data. Based on data, now think there is 95% probability that the true parameter is in the interval. BELIEF!

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

$$P(X = x|D) = \frac{P(X = x)P(D|X = x)}{P(D)} \tag{1}$$

$$P(X = x|D) = \frac{P(D|X = x)P(X = x)}{\sum(P(D|X = y)(P(X = y))} \tag{2}$$

$$P(X = x|D) \propto P(D|X = x)P(X = x) \tag{3}$$

statistical probability that a hypothesis is true, calculated in light of relevant observations. Relavant 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.

Bayesian Statistics

Contrast with Frequentist

In a frequentist framework a parameter of interest (mean, proportion, rate) is fixed and only varies due to sampling variation. The process of inferring the value of the 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 imaginary samples. However this is not possible many biological phenomia, where a hypothesis such as 'What is the porobabilty that a SNP is causal, given the data currently accrued?' Kirkwood2005 There is no interpretation of this hypothesis in frequentist terms, since there is a fixed but unknown quantity of the SNP (functional genomic issues such as gene expression under the frequentist interpretation). 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 current study, given the null hpyothesis; this framework does not allow for competing hypotheses. Lyle1999 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 being estimated. Bayesian statistics incorporate pre-exisitng 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 theorectical 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.

credible interval

A credible interval is a range of values within which an unobserved parameter values falls within a particular subjective probability. (Edwards, 1963) Unlike in frequentist statistics, in Bayesian we are asking what is the likelihood of a hypothesis, given the data we have observed. The data we have observed is represented by both priors - information that was already known that can be incorporated into analysis of the newly observed data.

credible set

A credible set is a set of objects, where it is believed one of the objects is the causal variant. Each object is a SNP, that has an assigned posterior probability of being the causal variant. The set is created by setting a threshold, where

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.

Link with Frequentist

Link between a frequentist confidence interval and a bayesian credible interval with an uninformative prior. When Bayesian framework is used, but the prior belief is uninformative, the posterior probability is derived from the observed data, defined by the likelihood. 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. Under this framework, a parameter of interest credible interval, the posterior probability a SNP is causal can be interpreted

Issues with Credible Sets

There is currently no definition for a credible set,

Scope of the Study:

The scope of the simulations and use cases in this analysis done under the assumptions that there is only one causal variant per credible set referred to as the single causal variant assumption. These limitations 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 and then integrating over those two posterior distributions to calculate probabilities that those

variants are shared Wallace2016. Therefore, this analysis may be able to be extended to involve colocalization in the future but is not addressed at this time.

Call 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 Gymrek2015. 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.

Finally this study does not address any trans-ancestry differences or geographical analysis of subpopulations and their respective potential subtypes. Different trans-ancestries have different LD patterns. A SNP that has a small p-value across groups that have different genetic architecture, shows a stronger genetic assocation 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.

Aims

This study aims to provide a history of fine mapping and its methods

entropy

A measure of a system's disorder. -sum p log(p)

Methods

How you did simulations

how you evaluated coverage

how you fixed (if we do!)

Results

plots of coverage vs size for unordered 'credible sets'

plots of coverage vs size

plots of entropy vs coverage-size

fix!

Discussion

References