UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Wits Bioinformatics

Introduction to Fine Mapping

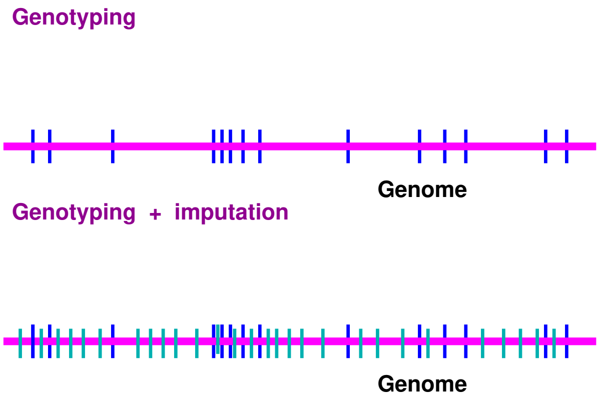
Scott Hazelhurst
Sydney Brenner Institute for Molecular Bioscience
School of Electrical & Information Engineering
University of the Witwatersrand
Johannesburg

2023

1 Introduction

- We are doing genotyping – know with *very* high confidence genotypes at ≈0.1% of genome.
- We may do imputation – statistically infer intermediate results using population LD information – know with *reasonable* confidence genotypes at ≈0.5% of genome

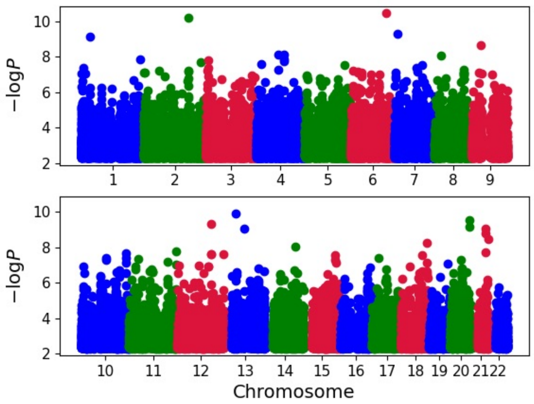
NB: The causal variant may not be captured and even if it is, may not have the lowest p value due to vagaries of noise in the data, especially if sample size is small.



Motivation

GWAS finds possible SNPs across the genome

- Pick your p, get out a set of associated SNPs



Association is not causality

Linkage Disequilibrium

- We expect real discoveries to have multiple SNPs
  - that have good p scores – above or close to cut-off ;
  - close to each other
- Which of the SNPs that meet a p-value cut-off are causal?

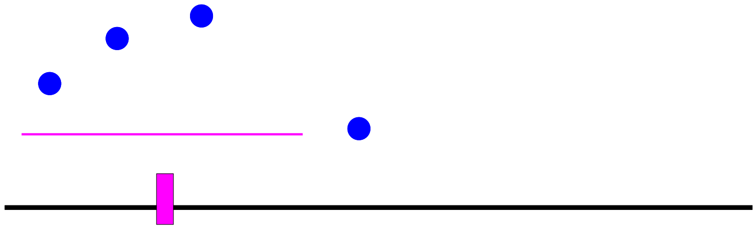
Association is not causality

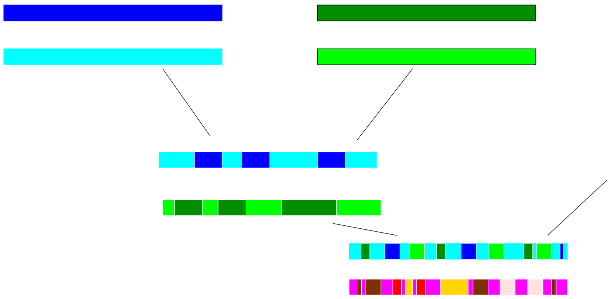
Causality requires showing

- that particular alleles at a SNP influence the biology
- goal – understand biological mechanisms from variation in genotype to variation in phenotype

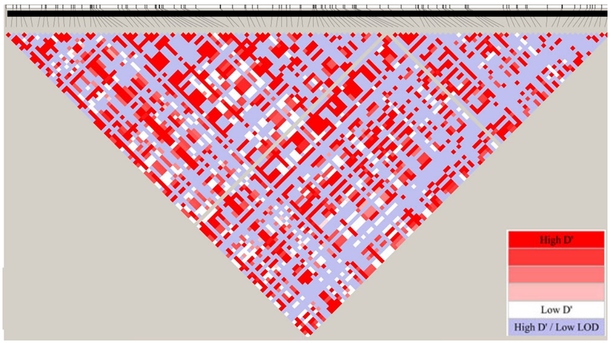
How done *in silico*?

- Using functional information (context, cell, disease, . . .)
- Using Linkage Disequilibrium



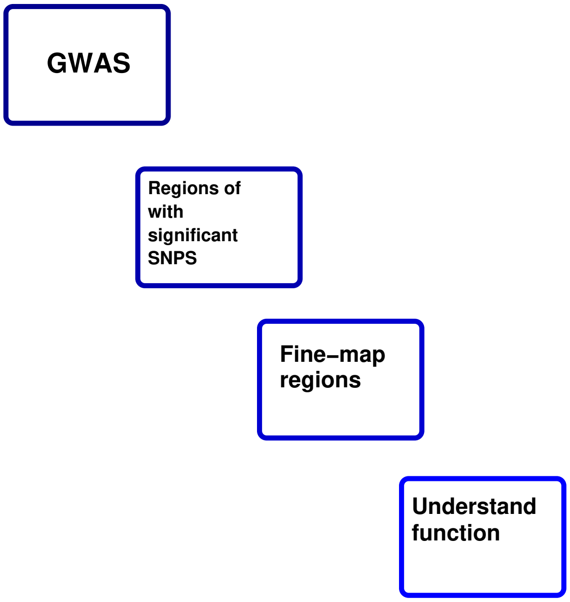


LD-blocks at population level



2 FM: Overview

Fine-mapping: Statistical approaches



GWAS

Regions of with significant SNPS

Fine−map regions

Understand function

Adapted from `doi.org/10.3389/fgene.2019.01304`

Basic input to process is results file – text file

- Chromosome, base position
- SNP ID (maybe)

- p-value
- β or odds ratio

May be sorted by p or position depending on purpose

- human readable in principle but . . .

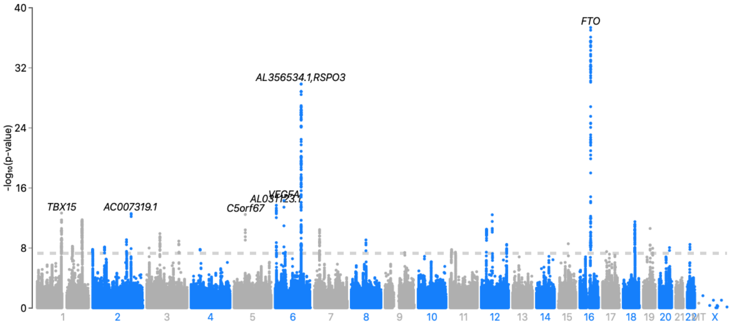
Takes input of results file

- Produces a Manhatten plot

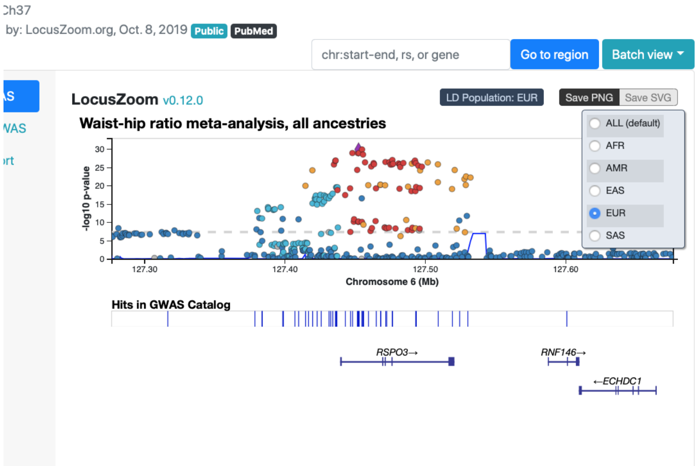
  `locuszoom.org`

Try: `https://my.locuszoom.org/gwas/826670/`

Study on waist-hip ratio



You can then click on a region.


Ch37

by: LocusZoom.org, Oct. 8, 2019

Waist-hip ratio meta-analysis, all ancestries

3 Starting the search. . .

Hang, on how do we know what the index SNP is? Important to remember what we are capturing in a GWAS.

Index or Lead SNP

- SNP with strongest evidence that it is associated
- May or may not be causal

Common that there are multiple index SNPs independent of each other.

- Need to distinguish SNPs in high LD with each other may all be in high LD with one single causal SNP

  Important issues

- Linkage disequilibrium patterns
- Density of SNP array
- Imputation and imputation quality

Determining the index SNP

Heuristic method – pick SNP with smallest p

Heuristic method + use of LD structure – pick SNP with smallest p

- Any other SNP in LD is assumed *not* to be causal.

3.1 Regression approaches

Forward/step-wise regression – test for independence

- Find the associated SNPs – order by p

  $\{s_0, s_1, \ldots s_{n-1}\}$
- Remove s_0
- Test $\{s_1, \ldots s_{n-1}\}$ using s_0 as a co-variate
- Any SNPs still associated?
- Repeat . . .

Forward regression is also called *conditioning*. We *condition* subsequent results based on assuming the lead SNP is a co-variate. Step-wise or forward regresstion can be expensive – you need to repeat many times if the credible set.

More seriously you lose power as you do this as we may not pick up independent signals – we can relax the p-cut-off but this may also mean we get false positives

Penalised regression models – do multiple regression with set of associated SNPs

Piece-wise regression For each SNP i independently solve

- $y_j = \beta_i x_{ij} + \epsilon_i$

Multiple regression Solve:

- $y_j = \sum(\beta_i x_{ij} + \epsilon_i)$

  Penalised regression : Add additional constraints to reduce impact of many of the SNPs

- e.g. Subject to $\sum \mid \beta_i \mid < \lambda$
- Goal : only those β_i that have biggest impact are picked
- Many β_i go to zero – those are *not* independent

Multiple methods : lasso, ElasticNet etc

3.2 Bayesian approaches

Bayesian models
Given m significant SNPs in a GWAS – which SNPs are *causal*?

- Model : set of SNPs that we *hypothesise* are causal

  e.g SNP 3, 33, 34, 37, 421 are causal – the rest are not

- Given m SNPs,
  - we represent a model as vector of m 0s and 1s
    e.g. $100110\ldots$ SNPs 0, 3, 4 are causal; 1 ,2, 5 are not.
  - So there are 2^m possible models

$P(M|D)$ – probability of the model given the data we see

- Goal: find the M with greatest probability

Use Bayes's rule:

- We can compute $P(D|M)$
- Apply Bayes:

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}$$

Mathematics of this are beyond the scope of this lecture. Typically we can (relatively) easily compute $P(D|M)$ – that is given the model we can find the probability that we see the data. Our model tells us which are the SNPs that are putatively causl and so from the GWAS stats we can work out P – the data being the set of probabilities and effects. We can then use a generalisation of Bayes's rule in order to compute the probability of the model

Bayes II: Posterior inclusion probability
PIP of SNP i: probability SNP is causal:

$$\mathit{PIP}_i = \sum\{P(M|D) : \text{SNP } i \text{ causal in } M\}$$

- May not be informative when several SNPs in high LD with each other and the causal SNP

Finding credible set formally – Bayesian method
Credible set: A set of SNPs that we think includes the causal SNP.

For model with one causal SNP – to be sure with probability α you have the causal SNP

- Test and find associated SNPs
- Rank SNPs by PIP (descending)
- Pick the smallest number so that sum of PIPs = α

Can be generalised to multiple causal SNPs

4 Multi-ancestry analyses

Meta-analyses

Recap

- Combining results from several different GWASes
- Effective way of increasing the sample size – improves power of analysis

General point: meta-analysis can combine results from different studies **using summary statistics only**

Multi-ancestry fine-mapping

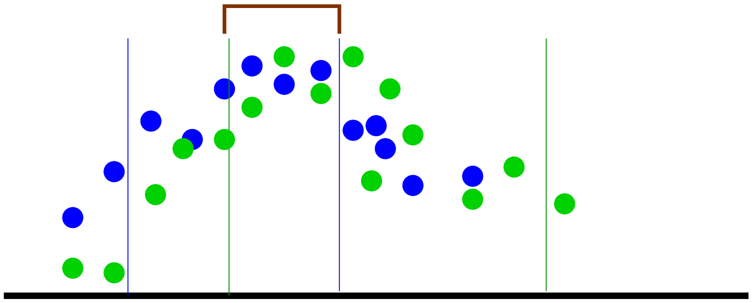
(aka, trans-ethnic mapping)

Doing meta-analysis using both summary statistics *and* LD-structure of the populations.

- If the studies come from different populations with different LD, then promotes fine-mapping
- African populations have smaller LD-blocks

The basic approach for two populations is

- do GWAS in both – find candidate SNPs in LD-blocks
- Use differences in LD-blocks to narrow range
- Combine summary statistics to compute new p values, β or ORs



5 Using biological information

Can we use biological annotation to narrow down search

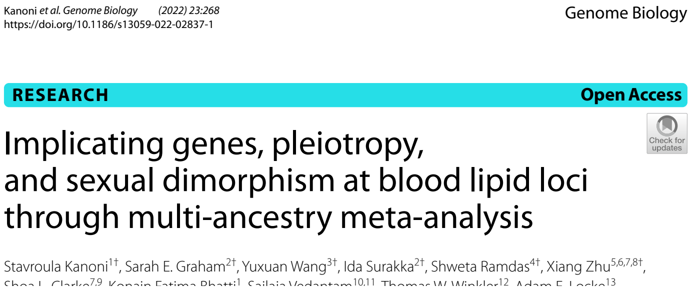
- Causality more plausible if there's a known biological function
- Guide follow up studies – prioritise
- Larger candidate sets make reduce the effectiveness and power of fine-mapping.
- Computational costs (less important issue)

Provided we are sure that we've tagged the region we may want to remove SNPs

Sources of information

- SNPs known to be causal or implicated already
- Genes known to be implicated already, gene annotation
- Coding regions – perhaps also using prediction of impact
- Splicing sites
- Regulatory regions, TSS, TFBS
- Functional elements (enhancers, promoters), tissue-specific effects

Case study



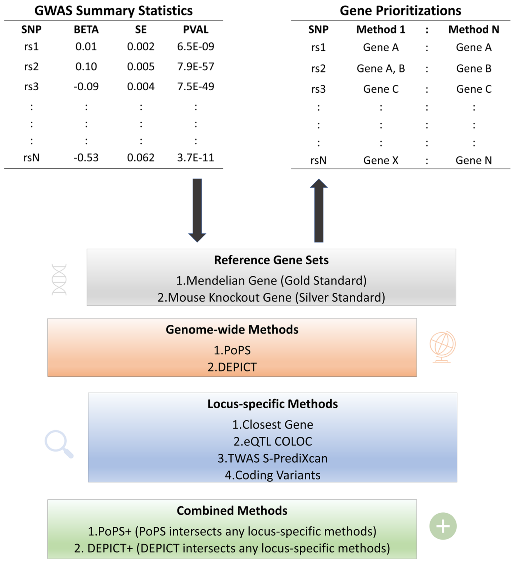
Kanoni *et al. Genome Biology* (2022) 23:268
https://doi.org/10.1186/s13059-022-02837-1

Genome Biology

RESEARCH Open Access

Implicating genes, pleiotropy, and sexual dimorphism at blood lipid loci through multi-ancestry meta-analysis

Stavroula Kanoni[1†], Sarah E. Graham[2†], Yuxuan Wang[3†], Ida Surakka[2†], Shweta Ramdas[4†], Xiang Zhu[5,6,7,8†],
Shoa L. Clarke[7,9], Konain Fatima Bhatti[1], Sailaja Vedantam[10,11], Thomas W. Winkler[12], Adam E. Locke[13]



GWAS Summary Statistics

SNP	BETA	SE	PVAL
rs1	0.01	0.002	6.5E-09
rs2	0.10	0.005	7.9E-57
rs3	-0.09	0.004	7.5E-49
:	:	:	:
:	:	:	:
:	:	:	:
rsN	-0.53	0.062	3.7E-11

Gene Prioritizations

SNP	Method 1	:	Method N
rs1	Gene A	:	Gene A
rs2	Gene A, B	:	Gene B
rs3	Gene C	:	Gene C
:	:	:	:
:	:	:	:
:	:	:	:
rsN	Gene X	:	Gene N

Reference Gene Sets
1.Mendelian Gene (Gold Standard)
2.Mouse Knockout Gene (Silver Standard)
Genome-wide Methods
1.PoPS
2.DEPICT
Locus-specific Methods
1.Closest Gene
2.eQTL COLOC
3.TWAS S-PrediXcan
4.Coding Variants
Combined Methods
1.PoPS+ (PoPS intersects any locus-specific methods)
2. DEPICT+ (DEPICT intersects any locus-specific methods)

6 References

References

- Broekema el al 2020. 'A practical view of fine-mapping and gene prioritization in the post-genome-wide association era.' *Open Biology*. **P10**(1). `https://doi.org/10.1098/rsob.190221`
- Kanoni et al (2022). Implicating genes, pleiotropy, and sexual dimorphism at blood lipid loci through multi-ancestry meta-analysis. *Genome Biology* **23**(268) `https://doi.org/10.1186/s13059-022-02837-1`
- Schaid et al 2018. 'From genome-wide associations to candidate causal variants by statistical fine-mapping.' *Nature Reviews Genetics* **19**(8) `https://rdcu.be/b8ZzW`